Structural Break and Time Series Modeling of Weekly Cereal Sales

Alex Spigler

April 4, 2025

Introduction

This report analyzes weekly cereal sales for a brand over a two-year period. At week 88, a competitor introduced a similar product into the market. The goal is to evaluate whether this competitive entry resulted in a structural change in sales patterns. Understanding structural breaks is important for businesses to identify the impact of interventions such as marketing campaigns, pricing changes, supply chain disruptions, or competitive actions.

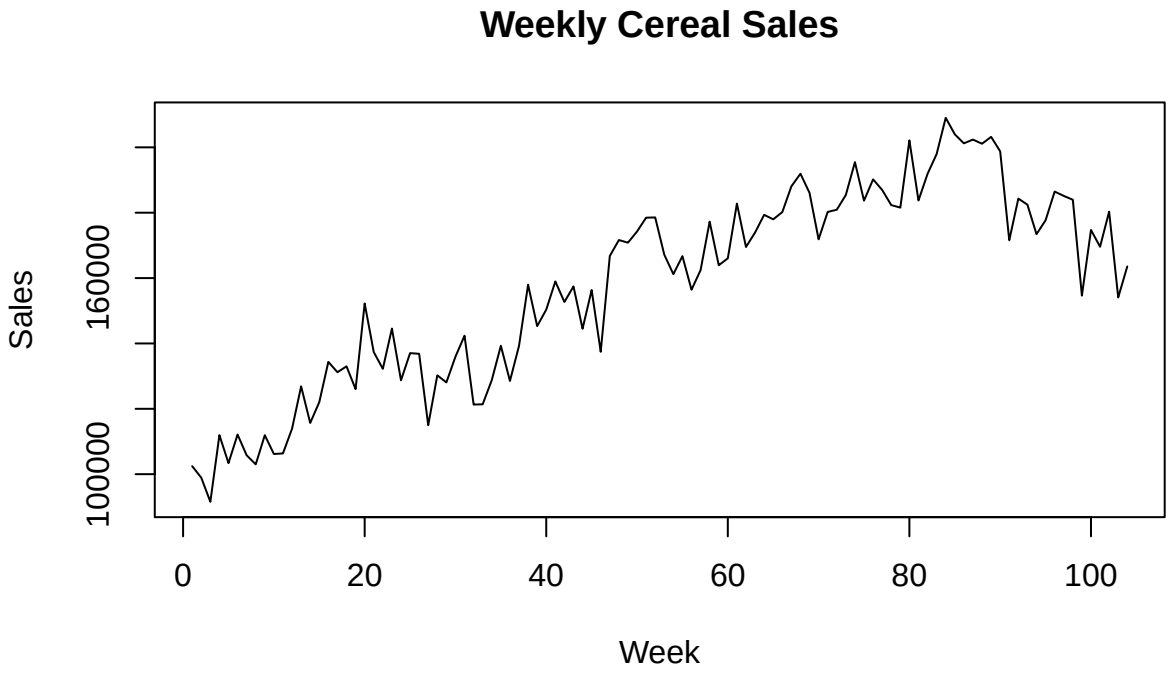
The goals of this analysis are to:

1. Model the underlying trend and a possible level/trend shift at the intervention week.
2. Diagnose any violations of classical regression assumptions, especially autocorrelation.
3. Produce short-term forecasts with prediction intervals.

The analysis is carried out in **R** using a regression model with an intervention term and ARIMA errors.

Data and Exploratory Analysis

```
cereal <- read.csv("cereal-sales-data.csv")
cereal_ts <- ts(cereal$sales, start = 1, frequency = 52)

# Plot with week numbers on x-axis
plot(
  x = 1:length(cereal_ts),
  y = as.numeric(cereal_ts),
  type = "l",
  main = "Weekly Cereal Sales",
  xlab = "Week",
  ylab = "Sales"
)
```



Visually, the series shows a non-constant mean and an upward trend, indicating nonstationarity. There also appears to be a noticeable change around week 88, motivating an intervention model.

Methods

Regression with Intervention

An intervention regression model is appropriate for detecting structural breaks. Let:

- y_t = cereal sales in week t
- $x_t = 0$ for $t < 88$, $x_t = 1$ for $t \geq 88$
- t = week index
- $x_t t$ = interaction term

The model is:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 t + \beta_3 x_t t + \varepsilon_t$$

This specification is appropriate because:

- β_0 captures the intercept (baseline level of sales) before week 88
- $\beta_2 t$ captures the trend (rate of change in sales) before week 88
- $\beta_1 x_t$ allows for a jump or drop in level at week 88
- $\beta_3 x_t t$ allows for a change in trend after week 88 (sales could start declining faster or growing more slowly)

```
n <- nrow(cereal)
cereal$t  <- seq_len(n)
cereal$x  <- ifelse(cereal$t < 88, 0, 1)
cereal$xt <- cereal$x * cereal$t
```

```r
reg_model <- lm(sales ~ x + t + xt, data = cereal)
summary(reg_model)
```

```
##
## Call:
## lm(formula = sales ~ x + t + xt, data = cereal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19259.9  -5478.9   -187.6   7303.8  26571.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 102585.02    2001.72  51.248  < 2e-16 ***
## x           278482.63   44089.18   6.316 7.46e-09 ***
## t             1153.18      39.51  29.186  < 2e-16 ***
## xt           -3256.83     459.89  -7.082 2.02e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9255 on 100 degrees of freedom
## Multiple R-squared:  0.9077, Adjusted R-squared:  0.9049
## F-statistic: 327.8 on 3 and 100 DF,  p-value: < 2.2e-16
```

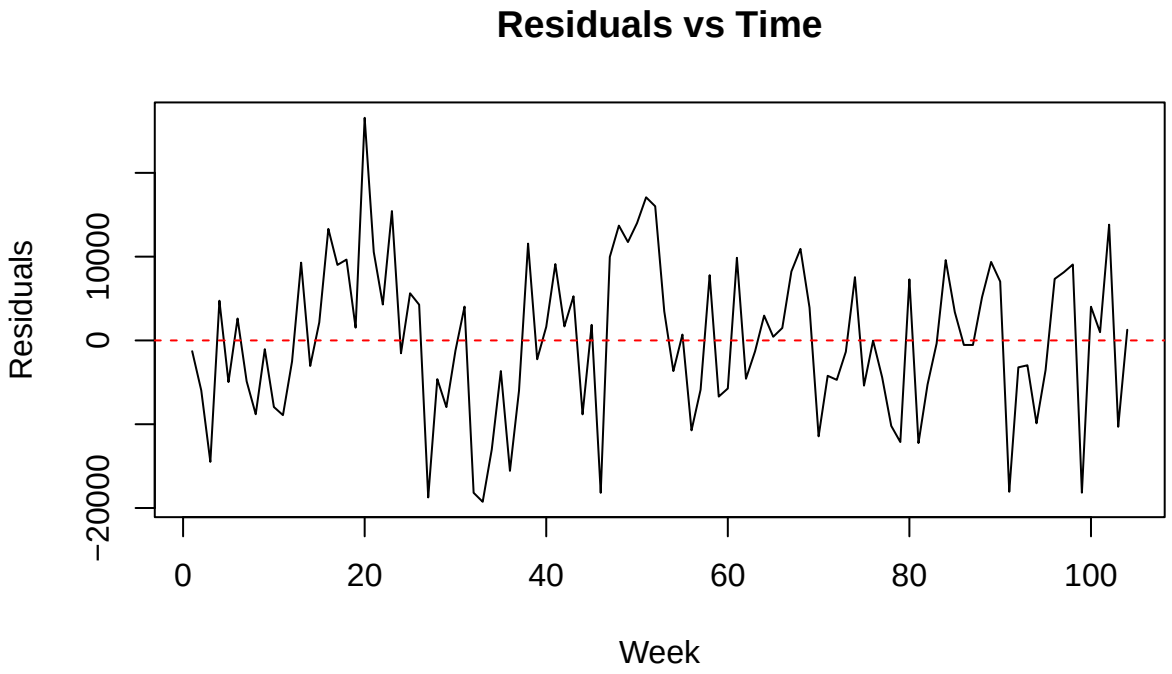
Interpretation of coefficients:

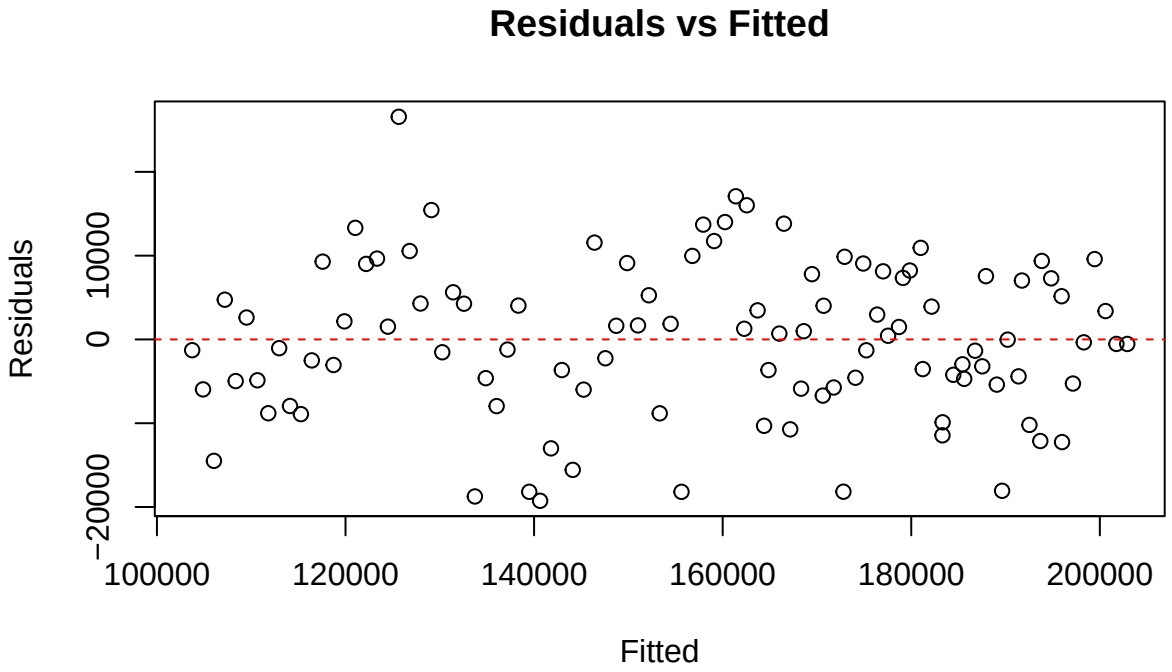
All coefficients are statistically significant ($p < 0.001$):

- $\hat{\beta}_0 = 102,585$: Baseline sales level before week 88
- $\hat{\beta}_2 = 1,153$: Pre-intervention trend (sales increased by approximately 1,153 units per week)
- $\hat{\beta}_1 = 278,483$: Immediate level shift at week 88 coinciding with competitor entry (a large jump in sales, possibly due to increased category awareness)
- $\hat{\beta}_3 = -3,257$: Change in slope after week 88 (sales declined by approximately 3,257 units per week after the intervention, resulting in a net negative trend post-intervention)

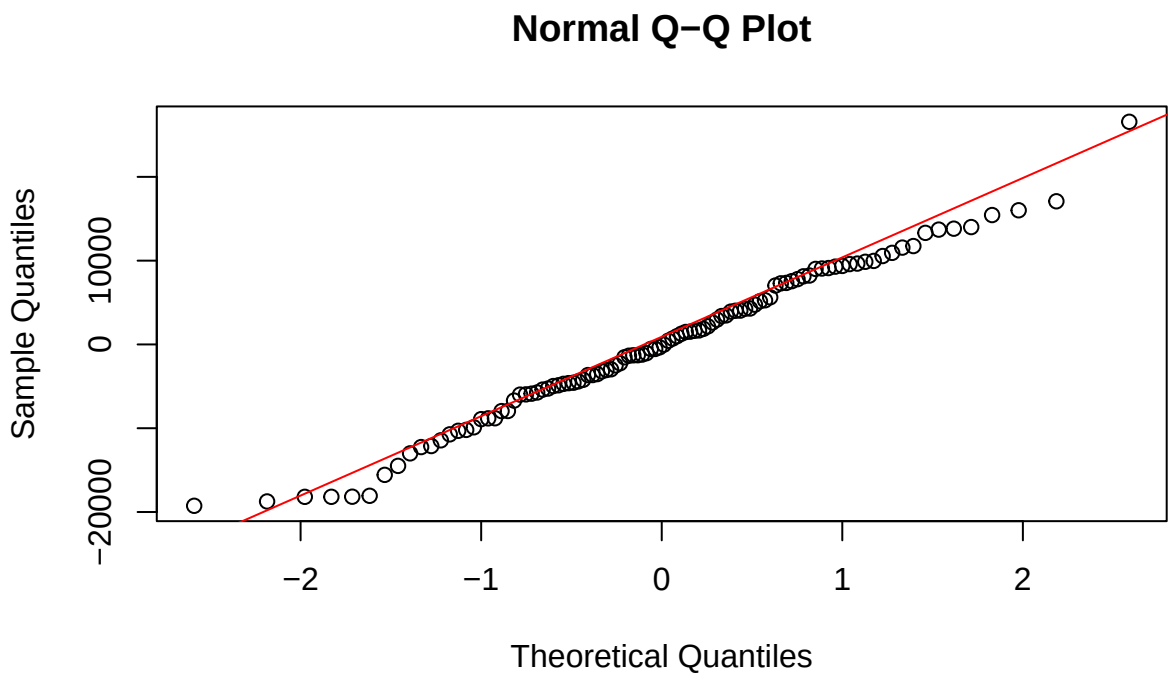
Regression Diagnostics

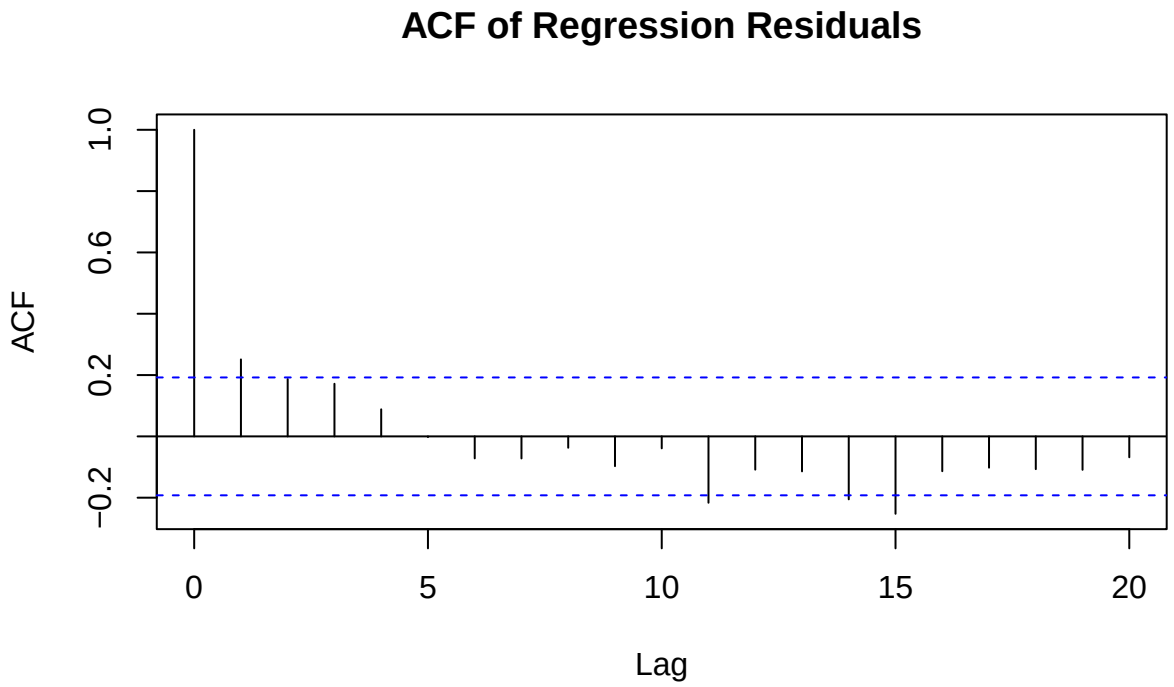
```r
reg_resid <- resid(reg_model)
par(mfrow = c(2, 1))

plot(reg_resid, type="l", main="Residuals vs Time", xlab="Week", ylab="Residuals")
abline(h=0, col="red", lty=2)

plot(fitted(reg_model), reg_resid, main="Residuals vs Fitted", xlab="Fitted", ylab="Residuals")
abline(h=0, col="red", lty=2)
```






```
qqnorm(reg_resid, main="Normal Q-Q Plot")
qqline(reg_resid, col="red")

acf(reg_resid, main="ACF of Regression Residuals")
```






```
par(mfrow = c(1, 1))
```

Tests

```
dwtest(reg_model)
```

```
##
```

```
##  Durbin-Watson test
##
## data:  reg_model
## DW = 1.4974, p-value = 0.001955
## alternative hypothesis: true autocorrelation is greater than 0
```

```
ncvTest(reg_model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.114541, Df = 1, p = 0.2911
```

```
shapiro.test(reg_resid)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reg_resid
## W = 0.98846, p-value = 0.5139
```

Independence: While the residuals vs. time plot does not show an obvious visual pattern, formal testing reveals positive autocorrelation. The Durbin-Watson test (p-value = 0.001955) rejects the null hypothesis of no autocorrelation. This is further confirmed by the ACF plot, which shows a significant spike at lag 1.

Constant Variance: The residuals vs. fitted values plot shows residuals randomly distributed around y = 0, suggesting constant variance. This is supported by the Non-Constant Variance Score Test (p-value = 0.2911), which does not reject the null hypothesis of constant variance.

Normality: The Q-Q plot appears approximately normal, and the Shapiro-Wilk test (p-value = 0.5139) does not reject the null hypothesis of normality.

Conclusion: The only assumption that fails is independence, due to the presence of positive autocorrelation among the error terms. This motivates the use of ARIMA errors.

Model Refinement with ARIMA Errors

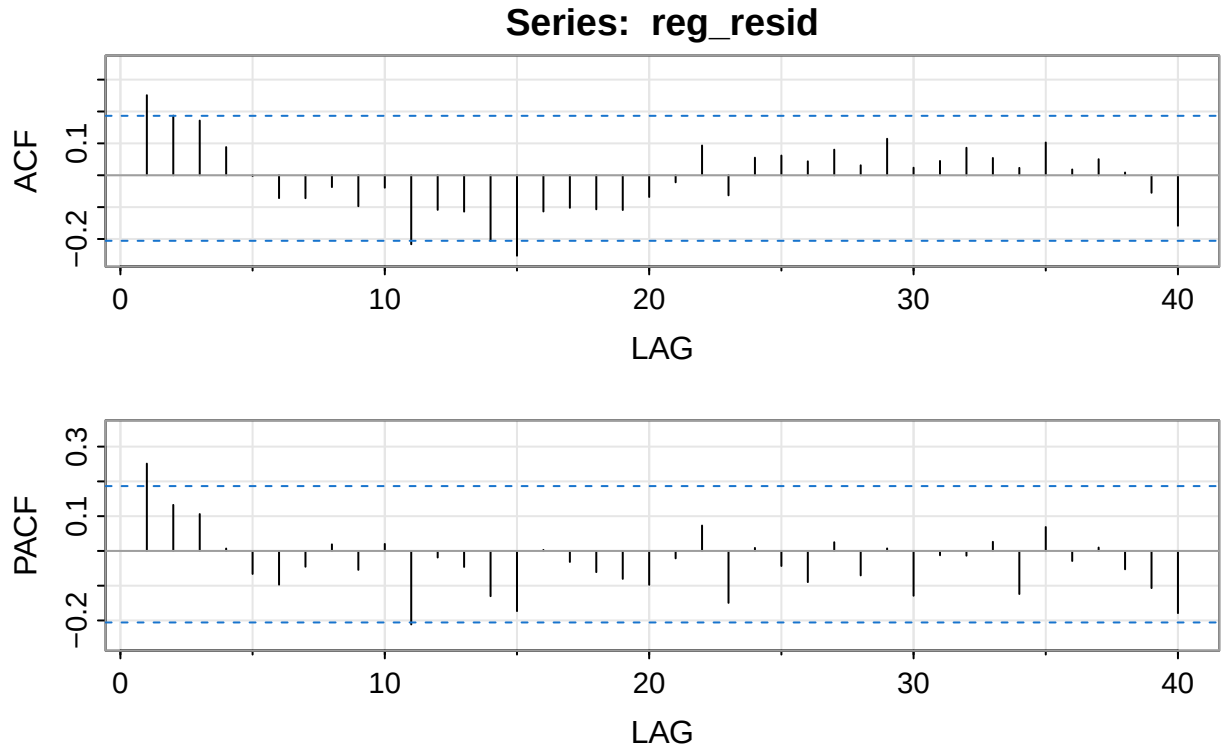
Residual Analysis

Because the diagnostic tests showed evidence of positive autocorrelation among the error terms, it is clear that the residuals are not white noise and require further modeling.

```
kpss.test(reg_resid)
```

```
##
##  KPSS Test for Level Stationarity
##
## data:  reg_resid
## KPSS Level = 0.047805, Truncation lag parameter = 4, p-value = 0.1
```

```
acf2(reg_resid, max.lag = 40)
```




```
##      [,1] [,2] [,3] [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11] [,12] [,13]
## ACF  0.25 0.19 0.17 0.09  0.00 -0.07 -0.07 -0.04 -0.10 -0.04 -0.22 -0.11 -0.11
## PACF 0.25 0.13 0.11 0.01 -0.07 -0.10 -0.05  0.02 -0.05  0.02 -0.21 -0.02 -0.05
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
## ACF  -0.21 -0.25 -0.11 -0.10 -0.11 -0.11 -0.07 -0.02  0.09 -0.06  0.06  0.06
## PACF -0.13 -0.17  0.00 -0.03 -0.06 -0.08 -0.10 -0.02  0.07 -0.15  0.01 -0.04
##      [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35] [,36] [,37]
## ACF   0.04  0.08  0.03  0.11  0.02  0.04  0.09  0.05  0.02  0.10  0.02  0.05
## PACF -0.09  0.02 -0.07  0.01 -0.13 -0.01 -0.01  0.03 -0.12  0.07 -0.03  0.01
##      [,38] [,39] [,40]
## ACF   0.01 -0.05 -0.16
## PACF -0.05 -0.11 -0.18
```

Stationarity Check: The KPSS test (p-value > 0.1) does not reject the null hypothesis of level stationarity, confirming the residuals have constant mean and variance. Therefore, differencing is not needed (d = 0 in ARIMA notation).

Seasonality Check: The ACF shows no significant spikes at seasonal lags such as 52 or 104, indicating no seasonal autocorrelation. Seasonal differencing is not appropriate for this series.

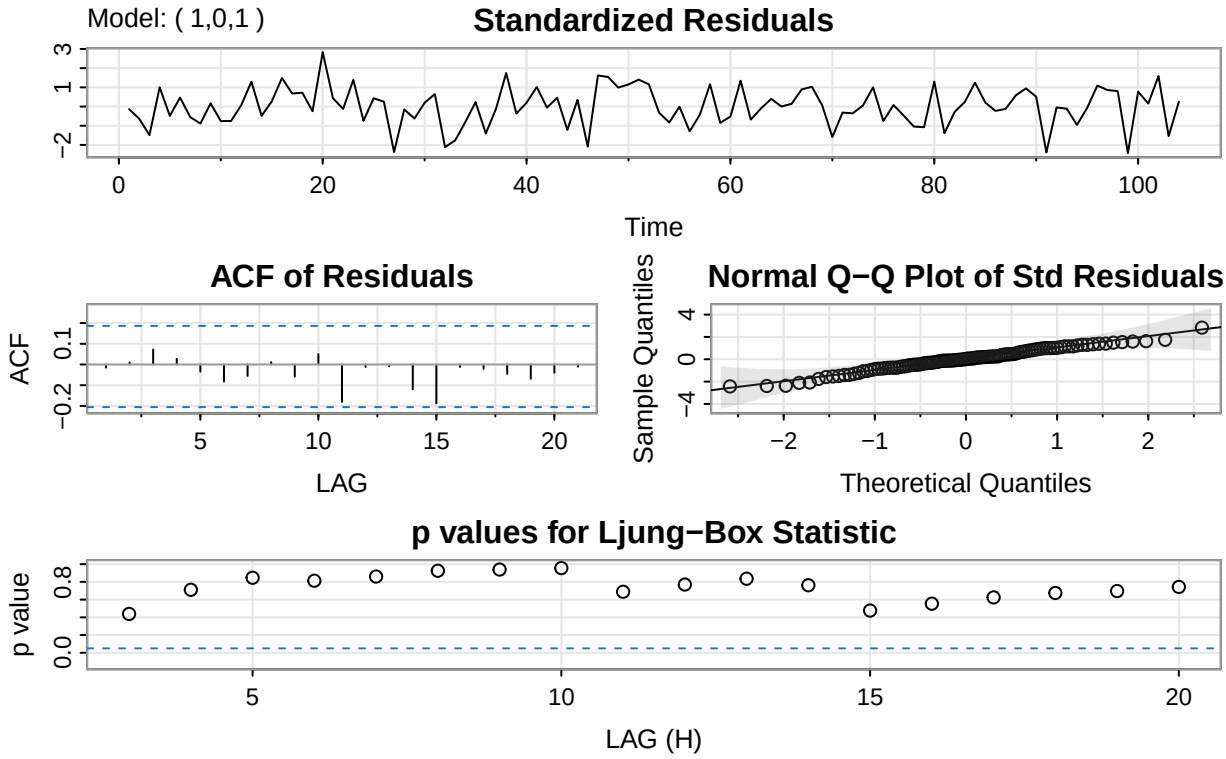
Model Selection: The ACF shows a significant spike at lag 1 that gradually tails off. The PACF similarly shows a significant spike at lag 1 that tails off. This pattern (both ACF and PACF showing initial spikes that decay) suggests an ARIMA(1,0,1) model is appropriate for the residuals.

ARIMA(1,0,1) Error Model

$$e_t = \phi_1 e_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

```r
sarima(reg_resid, 1, 0, 1)
```

```
## initial  value 9.118047
## iter   2 value 9.098328
## iter   3 value 9.088789
## iter   4 value 9.088293
## iter   5 value 9.087804
## iter   6 value 9.081453
## iter   7 value 9.075473
## iter   8 value 9.074289
## iter   9 value 9.073717
## iter  10 value 9.073354
## iter  11 value 9.073352
## iter  11 value 9.073352
## final  value 9.073352
## converged
## initial  value 9.069797
## iter   2 value 9.069757
## iter   3 value 9.069732
## iter   4 value 9.069729
## iter   5 value 9.069728
## iter   6 value 9.069725
## iter   7 value 9.069724
## iter   8 value 9.069723
## iter   8 value 9.069723
## final  value 9.069723
## converged
## <><><><><><><><><><><><><><><><><><><><>
##
## Coefficients:
##       Estimate        SE t.value p.value
## ar1     0.6728    0.1747  3.8513  0.0002
## ma1    -0.4514    0.2034 -2.2200  0.0287
## xmean -45.5615 1411.7509 -0.0323  0.9743
##
## sigma^2 estimated as 75407844 on 101 degrees of freedom
##
## AIC = 21.05425  AICc = 21.05655  BIC = 21.15595
##
```



Model Diagnostics:

- The plot of standardized residuals looks good and is randomized around y = 0
- The ACF of residuals is within bounds, indicating white noise
- The Q-Q plot appears normal
- The p-values for the Ljung-Box statistic are all above 0.05, indicating no remaining autocorrelation
- Both AR(1) and MA(1) terms are statistically significant (p-values of 0.0002 and 0.0287 respectively)
- AIC = 21.05425, BIC = 21.15595

This is a well-fitting model. Residuals now appear to be white noise, validating the combined regression + ARIMA(1,0,1) approach.

Combined Model Specification

The final model has a regression component with an intervention at Week 88, and ARIMA(1,0,1) errors to account for autocorrelation in the residuals:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 t + \beta_3 x_t t + e_t$$

where the error terms e_t follow an ARIMA(1,0,1) process:

$$e_t = \phi_1 e_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$$

with:

- ε_t = white noise
- $\phi_1 = 0.6728$ = autoregressive parameter
- $\theta_1 = -0.4514$ = moving average parameter

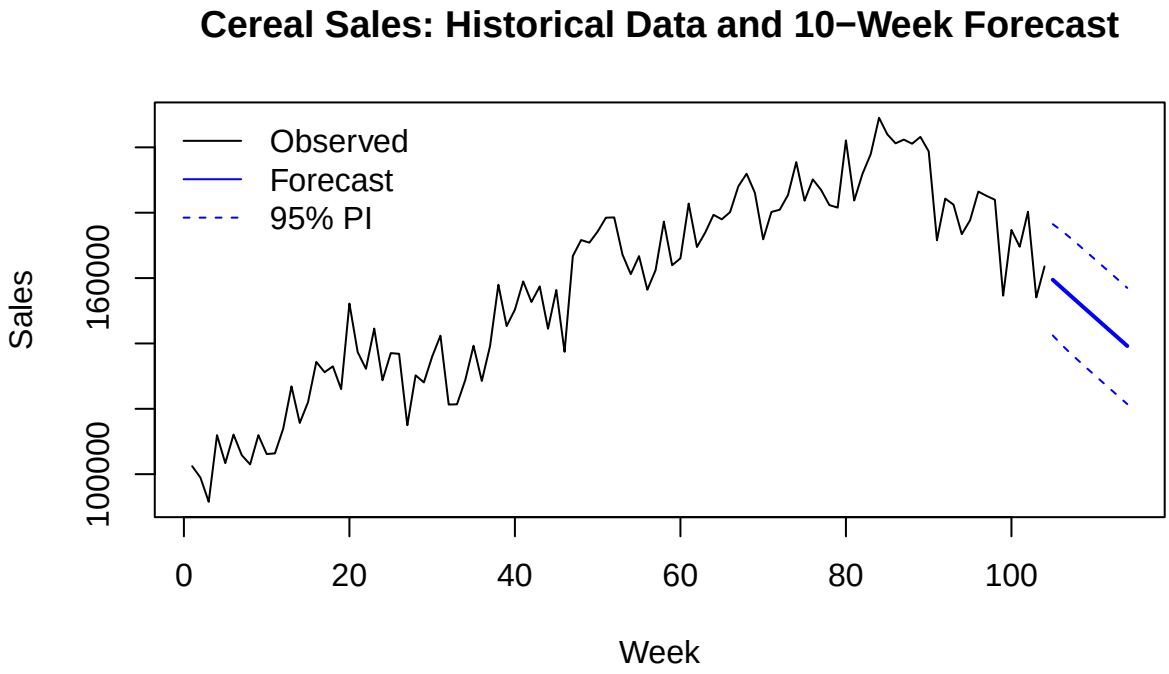
This model allows for a structural break in both the level and trend of sales at Week 88, and it captures autocorrelated, stationary errors.

Forecasting

```
xreg <- cbind(cereal$x, cereal$t, cereal$xt)

h <- 10
future_t  <- (n + 1):(n + h)
future_x  <- ifelse(future_t < 88, 0, 1)
future_xt <- future_x * future_t
newxreg <- cbind(future_x, future_t, future_xt)

fc <- sarima.for(
  x       = cereal_ts,
  n.ahead = h,
  p       = 1,
  d       = 0,
  q       = 1,
  xreg    = xreg,
  newxreg = newxreg,
  plot    = FALSE
)

forecast_df <- data.frame(
  Week     = future_t,
  Forecast = round(fc$pred, 2),
  Lower95  = round(fc$pred - 1.96 * fc$se, 2),
  Upper95  = round(fc$pred + 1.96 * fc$se, 2)
)
forecast_df
```

```
##    Week Forecast  Lower95  Upper95
## 1   105 159470.0 142456.0 176483.9
## 2   106 157193.5 139761.0 174626.1
## 3   107 154929.9 137309.5 172550.3
## 4   108 152675.0 134969.4 170380.6
## 5   109 150426.0 132681.7 168170.3
## 6   110 148180.9 130418.9 165942.9
## 7   111 145938.6 128168.5 163708.6
## 8   112 143698.0 125924.3 161471.7
## 9   113 141458.7 123683.3 159234.1
## 10  114 139220.2 121444.0 156996.4
```

```r
# Base plot: historical data
plot(
  x = 1:length(cereal_ts),
  y = as.numeric(cereal_ts),
  type = "l",
  xlim = c(1, n + h),
  ylim = range(
    c(
      as.numeric(cereal_ts),
      fc$pred + 2 * fc$se,
      fc$pred - 2 * fc$se
    )
  ),
  main = "Cereal Sales: Historical Data and 10-Week Forecast",
  xlab = "Week",
  ylab = "Sales"
)

# Future x-values (weeks)
future_index <- (n + 1):(n + h)

# Forecast line
lines(future_index, fc$pred, col = "blue", lwd = 2)

# 95% prediction intervals
lines(future_index, fc$pred + 1.96 * fc$se, col = "blue", lty = 2)
lines(future_index, fc$pred - 1.96 * fc$se, col = "blue", lty = 2)

legend(
  "topleft",
  legend = c("Observed", "Forecast", "95% PI"),
  col    = c("black", "blue", "blue"),
  lty    = c(1, 1, 2),
  bty    = "n"
)
```



Conclusion

This analysis successfully modeled weekly cereal sales with a structural break at week 88 using a combined regression and ARIMA approach.

Key Findings:

- Regression analysis revealed a significant structural break at week 88, coinciding with the competitor's market entry. Results show an immediate jump in sales (+278,483 units, likely due to increased category awareness) followed by a declining trend (-3,257 units/week, consistent with competitive pressure).
- Standard regression residuals exhibited significant positive autocorrelation, violating the independence assumption.
- An ARIMA(1,0,1) error structure effectively captured the autocorrelation, with all diagnostic tests confirming the residuals now behave as white noise.
- The combined model fits the data well and produces stable 10-week forecasts with reasonable prediction intervals.

Potential Extensions:

Future work could explore:

- Analysis of competitor pricing, marketing spend, or product features to explain the magnitude of the sales impact
- Incorporating additional explanatory variables (such as promotions or other competitor actions)
- Extending the forecast horizon and validating predictions against actual data