Analyzing the Law of One Price through Banana Pricing Regression Models

Chris Grace

2025-12-05

Introduction: Law of One Price

In economics, the Law of One Price (LOOP) states that in the absence of trade barriers, with free competition, and price flexibility, the price of a good across markets is the same in a common currency.

To investigate the law, we'll consider a commodity good, **bananas**, and explore price discrepancy across markets. To attempt to disprove the LOOP, we'll build a regression model and attempt to predict banana prices. If there's enough price discrepancy between markets for us to construct a usable model, we'll have evidence against the LOOP.

Part 1: Data Collection

Before we go to the store and see how expensive bananas are, we should figure out what data we need to collect.

We definitely need the price of bananas in the local currency, but what do we define as "bananas"? For our analysis, let's consider 1KG of bananas as a bunch, and use the wholesale per-kilogram price of bananas as our price level. Wholesale prices cut out retail price volatility and most of the markup, so we'll get a more representative metric.

To compute a global price index, we'll also need the exchange rate data to relate the price in local currency to a standard currency. The US dollar is the clear choice here, so we'll collect the average spot exchange rate between local currency and USD for each country.

Most macroeconomic data for 2025 has not been finalized at the time of this analysis, so we'll collect data for 2024. To include as many countries as possible, we'll use a generative AI tool with web search capabilities to do our research for us. These tools are incredibly convenient but require a lot of oversight, both during prompting and when validating the dataset. To obtain the data, I used GPT 5's deep research mode with the following prompt:

> "For at least 50 countries providing a representative sample of the world, compile a .CSV format data set containing four columns:"country", "continent", "price", and "e", where "price" is the wholesale price/KG (in 2024 local currency) of a bunch of bananas and "e" is the spot exchange rate of the local currency with the US Dollar. Define a "bunch" of bananas as 1 kilogram of bananas for simplicity. Prioritize accuracy and provenance of the data, make necessary approximations or estimations but clearly note when you do so."

The model requested clarification on the desired exchange rate, source preferences, retail/wholesale prices, and the selection of countries to consider. I provided the following answers:

- The exchange rate should be the average for the entire 2024 year
- Prioritize institutional sources over firm-level data whenever possible, but retail data can be used in markets where wholesale prices are difficult to find
- Only include sovereign nations, consider a representative sample of countries across the globe.

For the predictive part of our analysis, we'll also need some auxiliary data points for each country. I'm choosing to include the following additional data points for each country:

- **Banana Production**: This seems like the greatest determinant of banana price in each country, so we'll include a term for banana production in tonnes per 1000 people.
- **Banana Imports**: Countries that import lots of bananas are more exposed to transaction costs and trade barriers that prevent the LOOP from holding. We can collect import data and investigate it's relationship with banana prices.
- **GDP Per Capita**: We expect countries with higher incomes to have higher prices for consumer goods, so GDP per capita is probably a good predictor of banana price

- **Logistics Performance Index (LPI)** : Another way of tracking the ease of foreign transactions is through the World Bank's Logistic Performance Index which measures the efficiency of logistics in each country. We expect higher logistic efficiency to imply lower banana prices.
- **Inflation**: Inflation data for the 2024 year, which may introduce price discrepancies for countries that wouldn't otherwise have them (eg. Argentina, Turkey)

Adding the above predictors to the prompt, we get figures for each feature with some estimation/imputation on the LLM's part. Thankfully this time there's a separate column for the "notes":

```
bananas_raw <- read.csv(file = "bananas.csv", header = TRUE)

# count the number of countries
nrow(bananas_raw)
```

```
## [1] 55
```

```
# get the column names
names(bananas_raw)
```

```
##  [1] "country"   "continent" "price"     "e"         "prod"      "imports"
##  [7] "percapgdp" "lpi"       "inflation" "notes"
```

```
# show a preview
head(bananas_raw)
```

```
##      country     continent  price      e  prod imports percapgdp  lpi inflation
## 1        USA North America   1.66   1.00   2.9     0.5     75994 3.89       3.4
## 2     Canada North America   1.46   1.37   0.9     0.2     53623 3.66       2.8
## 3     Mexico North America  17.00  18.33  14.2     1.8     10002 3.40       5.8
## 4  Guatemala North America   6.50   7.70  85.1     0.2      5230 2.50       4.6
## 5   Honduras North America  28.00  24.15  62.0     0.1      3270 2.40       5.0
## 6 Costa Rica North America 420.00 540.00 110.5     6.8     13120 2.85       4.2
##                                                                                notes
## 1                                                price: USDA market aggregates/Tridge 2024; e: USD
## 2                                    price: Canadian wholesale market reports 2024; e: IMF 2024 avg
## 3      price: Mexico central wholesale bulletin 2024; e: IMF 2024 avg; prod/imports: FAOSTAT (2023)
## 4                         price: national market bulletin/Tridge 2024; prod/imports: FAOSTAT (2023)
## 5                         price: national market bulletin/Tridge 2024; prod/imports: FAOSTAT (2023)
## 6 price: Tridge/exporter port price (USD/kg converted to CRC using e); prod/imports: FAOSTAT (2023)
```

To validate the accuracy of the data set (never trust AI!!), I performed a 1-in-11 systematic sample and manually verified the figures for each country using the model's data sources. While all of the information I verified was correct, I noticed a few misformatted exchange rates in the data set that I manually corrected before importing the data. As we continue our analysis, I'll note any inconsistencies and how I fixed them.

Gen AI statement:

AFTER DATA COLLECTION, GENERATIVE AI WAS NOT USED AT ANY POINT

Part 2: EDA and Visualization

Our first step to check for price discrepancy is to construct a global price index. The corollary to the LOOP, purchasing power parity, suggests that exchange rates move to "equalize" the price level across regions through arbitrage mechanisms.

PPP $\implies P_{US} = P_{FX} / E_{FX/US}$ Where sub FX denotes the price/exchange rate in a foreign country. Thus, we can compute the global price level by dividing the foreign price by the exchange rate.

- Correction Note: I noticed incorrect exchange rates for Papua New Guinea and Jamaica, so I used the 2024 average according to the IMF

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
bananas <- bananas_raw |>
           select(-notes) |>
           rename(price_raw = price) |>
           mutate(price = as.numeric(price_raw / e)) |>
           select(country, continent, price_raw, price, everything())

head(bananas)
```

```
##      country     continent price_raw     price      e  prod imports percapgdp
## 1        USA North America      1.66 1.6600000   1.00   2.9     0.5     75994
## 2     Canada North America      1.46 1.0656934   1.37   0.9     0.2     53623
## 3     Mexico North America     17.00 0.9274414  18.33  14.2     1.8     10002
## 4  Guatemala North America      6.50 0.8441558   7.70  85.1     0.2      5230
## 5   Honduras North America     28.00 1.1594203  24.15  62.0     0.1      3270
## 6 Costa Rica North America    420.00 0.7777778 540.00 110.5     6.8     13120
##    lpi inflation
## 1 3.89       3.4
## 2 3.66       2.8
## 3 3.40       5.8
## 4 2.50       4.6
## 5 2.40       5.0
## 6 2.85       4.2
```

With our price index complete, we can check the average price per continent:

```
cont_table <- bananas |>
              group_by(continent) |>
              summarize(mean_price = round(mean(price), 3)) |>
              arrange(desc(mean_price))

cont_table
```

```
## # A tibble: 7 × 2
##   continent     mean_price
##   <chr>              <dbl>
## 1 Europe             2.18
## 2 Oceania            1.45
## 3 North America      1.18
## 4 Africa             0.925
## 5 Asia               0.848
## 6 Europe/Asia        0.599
## 7 South America      0.431
```

Just looking at the price per continent, it seems highly likely that there are price discrepancies across the globe. We can get a better idea of the distribution of prices using a world map.

For ideal visualization, I'm making the scale follow banana colors :). Countries with higher prices will take greener hues while countries with lower prices are yellower.

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
library(ggspatial)

world_coords <- map_data("world") |>
                filter(region != "Antarctica")

world_plot <- ggplot() +
              geom_map(data = world_coords, map = world_coords,
                       aes(group = group, map_id = region),
                           fill = "white", color = "black", linewidth = 0.5) +
              geom_map(data = bananas, map = world_coords,
                       aes(fill = price, map_id = country),
                           color = "#5b5b5b", linewidth = 0.5) +
              coord_map("rectangular", lat0 = 0, xlim = c(-180,180), ylim = c(-60, 90)) +
              scale_fill_continuous(low = "#DFB210", high = "#35bf52", guide = "colorbar") +
              labs(fill = "Price (2024 USD)") +
              ggtitle("Worldwide Banana Prices",
                      subtitle = "Wholesale 2024 Price/KG for selected countries") +
              ylab("") + xlab("") +
              theme_minimal() +
              theme(panel.background = element_rect(fill = '#87B8DF', color = 'black'))

world_plot
```


```
# dev.copy(device = png, filename = "world_plot.png", width = 1500, height = 750)
# dev.off()
```

Our map provides a more useful gauge of price distribution across the world. The general trends are consistent with the table - prices are highest in Europe, North America, and Oceania, and lowest in South America and Asia. Among all the continents, it seems that Africa has the greatest variation in price. We can investigate the price variation across each continent by adding standard deviation to our previous table:

```
# previous table was already grouped so we need to remake it
cont_table <- bananas |>
              group_by(continent) |>
              summarize(mean_price = round(mean(price), 3),
                        std_dev = round(sd(price), 3)) |>
              arrange(desc(std_dev))

cont_table
```

```
## # A tibble: 7 × 3
##   continent     mean_price std_dev
##   <chr>              <dbl>   <dbl>
## 1 Asia               0.848   0.625
## 2 Africa             0.925   0.553
## 3 North America      1.18    0.5
## 4 Oceania            1.45    0.317
## 5 South America      0.431   0.275
## 6 Europe             2.18    0.269
## 7 Europe/Asia        0.599   0.241
```

Asia and Africa have the highest variation in banana prices, although North America is a close third place. Given the large variation in mean price country to country, it seems reasonable to include a variable for continent in our model. If this variable is significant, we have some evidence against the LOOP.

Now that we're familiar with the price distribution across each continent and the globe, we can begin investigating each of the other features we collected. We'll use scatter plots to check out the shape, strength, and direction of the relationship between each feature and price.

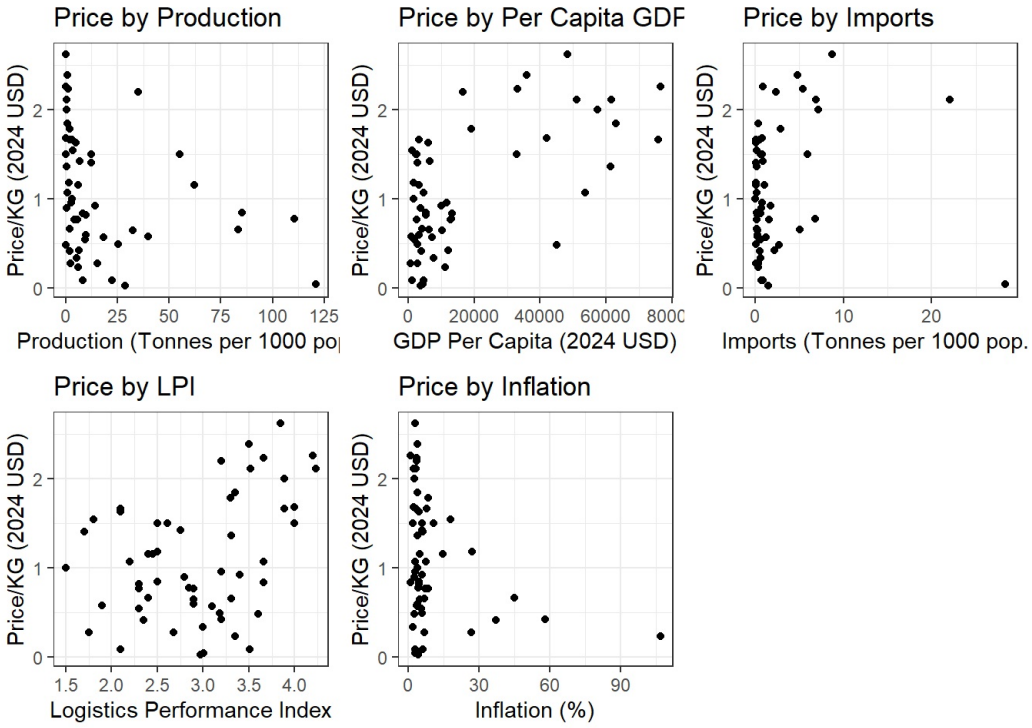
```
library(ggplot2)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.3.3
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
prod_plot <- ggplot(bananas, aes(x = prod, y = price)) +
             geom_point() + theme_bw() +
             xlab("Production (Tonnes per 1000 pop.)") +
             ylab("Price/KG (2024 USD)") +
             ggtitle("Price by Production")

gdp_plot <- ggplot(bananas, aes(x = percapgdp, y = price)) +
            geom_point() + theme_bw() +
            xlab("GDP Per Capita (2024 USD)") +
            ylab("Price/KG (2024 USD)") +
            ggtitle("Price by Per Capita GDP")

import_plot <- ggplot(bananas, aes(x = imports, y = price)) +
            geom_point() + theme_bw() +
            xlab("Imports (Tonnes per 1000 pop.)") +
            ylab("Price/KG (2024 USD)") +
            ggtitle("Price by Imports")

lpi_plot <- ggplot(bananas, aes(x = lpi, y = price)) +
            geom_point() + theme_bw() +
            xlab("Logistics Performance Index") +
            ylab("Price/KG (2024 USD)") +
            ggtitle("Price by LPI")

inflation_plot <- ggplot(bananas, aes(x = inflation, y = price)) +
                  geom_point() + theme_bw() +
                  xlab("Inflation (%)") +
                  ylab("Price/KG (2024 USD)") +
                  ggtitle("Price by Inflation")

grid.arrange(prod_plot, gdp_plot, import_plot, lpi_plot, inflation_plot,
             ncol = 3, nrow = 2)
```




```
# dev.copy(device = png, filename = "predictor_plots.png", width = 1000, height = 500)
# dev.off()
```

The ideal predictor has a high correlation with the response variable, so we're looking for a pattern similar to a straight line or otherwise identifiable trend. Apart from LPI, each predictor has a large number of countries clustered around the x = 0 line. These points will limit the usefulness of each predictor, but there may be relationships between variables that turn out to be useful. While I expect inflation and imports are too clustered to be significant predictors, we observe the hierarchy principle and leave them in the model. The hierarchy principle requires the parent predictors of interaction terms to remian in the model. Worst case, we have an insignficant predictor that explain miniscule amounts of price variation.

Unlike inflation and imports, LPI and per-capita GDP both appear to be moderately correlated with price. We can quantify this relationship:

```
# check correlation for LPI
cor(bananas$lpi, bananas$price)
```

```
## [1] 0.3683742
```

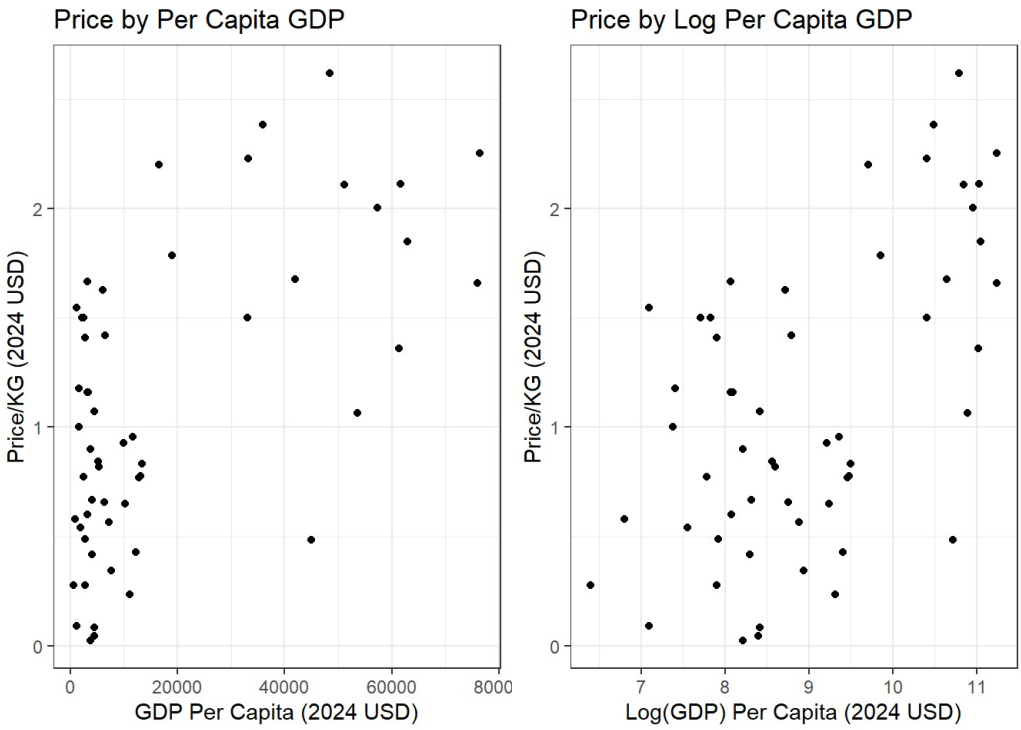
```
# check correlation for per-capita GDP
cor(bananas$percapgdp, bananas$price)
```

```
## [1] 0.607906
```

We can see that per-capita GDP is most strongly correlated with banana prices. High correlation is good, but the cluster of points around the Y-axis is still not ideal. From my economics classes I know that per capita GDP is left skewed, so a log transformation would make sense. We can apply the transformation and see if the new variable looks more linear.

I'm making another data frame to use for modeling

```
banana_modeling <- bananas |>
                   select(continent, price, percapgdp, imports, lpi, inflation) |>
                   mutate(log_percapgdp = log(percapgdp))

gdp_trans_plot <- ggplot(banana_modeling, aes(x = log_percapgdp, y = price)) +
                  geom_point() + theme_bw() +
                  xlab("Log(GDP) Per Capita (2024 USD)") +
                  ylab("Price/KG (2024 USD)") +
                  ggtitle("Price by Log Per Capita GDP")

grid.arrange(gdp_plot, gdp_trans_plot, nrow = 1, ncol = 2)
```



The graph on the right shows the transformed variable, which looks much more linear than before. With linear regression, we want our predictor-response relationships to be as linear as possible, so this transformation will help improve the accuracy of our models. Speaking of models:

Model Building

Now with a better understanding of our data, we can begin building a model to predict the price of bananas for a given country. We'll start with a naive baseline using all available predictors to test against:

```
naive_mod <- lm(price ~ ., banana_modeling)

summary(naive_mod)
```

```
##
## Call:
## lm(formula = price ~ ., data = banana_modeling)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82411 -0.22160 -0.02914  0.23589  1.01307
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -6.880e-01  1.086e+00  -0.633   0.5299
## continentAsia          -3.253e-01  2.210e-01  -1.472   0.1483
## continentEurope         6.569e-01  3.094e-01   2.123   0.0395 *
## continentEurope/Asia   -5.801e-01  3.971e-01  -1.461   0.1513
## continentNorth America -1.625e-01  2.415e-01  -0.673   0.5045
## continentOceania        2.927e-02  3.019e-01   0.097   0.9232
## continentSouth America -5.537e-01  3.092e-01  -1.791   0.0804 .
## percapgdp               3.057e-06  7.336e-06   0.417   0.6790
## imports                -1.099e-02  1.610e-02  -0.683   0.4984
## lpi                    -1.874e-01  2.756e-01  -0.680   0.5002
## inflation              -3.933e-03  4.116e-03  -0.955   0.3447
## log_percapgdp           2.705e-01  1.796e-01   1.506   0.1393
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4533 on 43 degrees of freedom
## Multiple R-squared:  0.6442, Adjusted R-squared:  0.5531
## F-statistic: 7.077 on 11 and 43 DF,  p-value: 1.125e-06
```

Our naive baseline model explains 55.314% of variation in banana prices. From the output, we can see that only the indicator variables for Europe and South America provide significant explanatory power on their own. Let's test a model that includes interaction terms between predictors:

```
mod2 <- lm(price ~ .^2, banana_modeling)

summary(mod2)
```

```
##
## Call:
## lm(formula = price ~ .^2, data = banana_modeling)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.39819 -0.01117  0.00000  0.03492  0.48658
##
## Coefficients: (6 not defined because of singularities)
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -3.350e+01  1.959e+01  -1.710  0.12144
## continentAsia                     -5.245e+00  1.179e+01  -0.445  0.66690
## continentEurope                    2.037e+02  1.311e+02   1.555  0.15447
## continentEurope/Asia               6.968e+01  1.643e+02   0.424  0.68144
## continentNorth America             2.708e+00  1.514e+01   0.179  0.86203
## continentOceania                  -1.054e+00  2.659e+00  -0.397  0.70088
## continentSouth America             6.128e+01  3.770e+01   1.626  0.13847
## percapgdp                          1.343e-03  2.165e-03   0.620  0.55053
## imports                           -1.650e+01  6.085e+00  -2.711  0.02395 *
## lpi                                1.544e+01  1.468e+01   1.052  0.32042
## inflation                          2.856e-01  3.464e-01   0.824  0.43103
## log_percapgdp                      4.937e+00  2.281e+00   2.165  0.05861 .
## continentAsia:percapgdp            5.480e-05  3.257e-04   0.168  0.87012
## continentEurope:percapgdp          7.535e-04  4.138e-04   1.821  0.10199
## continentEurope/Asia:percapgdp    -5.532e-03  1.286e-02  -0.430  0.67709
## continentNorth America:percapgdp   1.510e-04  2.745e-04   0.550  0.59579
## continentOceania:percapgdp         2.933e-04  2.363e-04   1.241  0.24595
## continentSouth America:percapgdp  -9.294e-04  7.926e-04  -1.173  0.27108
## continentAsia:imports             -1.634e+00  5.027e-01  -3.251  0.00998 **
## continentEurope:imports           -2.834e+00  1.399e+00  -2.026  0.07340 .
## continentEurope/Asia:imports              NA         NA      NA       NA
## continentNorth America:imports    -2.552e+00  1.021e+00  -2.500  0.03387 *
## continentOceania:imports          -1.588e+00  3.706e+00  -0.429  0.67831
## continentSouth America:imports    -1.721e+00  6.421e-01  -2.681  0.02518 *
## continentAsia:lpi                  7.680e-01  7.032e-01   1.092  0.30317
## continentEurope:lpi               -1.502e+01  6.281e+00  -2.392  0.04044 *
## continentEurope/Asia:lpi                  NA         NA      NA       NA
## continentNorth America:lpi         2.322e+00  1.389e+00   1.672  0.12895
## continentOceania:lpi              -1.058e-01  1.488e+00  -0.071  0.94486
## continentSouth America:lpi        -1.759e+01  1.042e+01  -1.689  0.12556
## continentAsia:inflation            2.623e-02  2.161e-02   1.214  0.25575
## continentEurope:inflation         -6.266e-01  6.272e-01  -0.999  0.34390
## continentEurope/Asia:inflation            NA         NA      NA       NA
## continentNorth America:inflation  -4.828e-01  6.408e-01  -0.753  0.47047
## continentOceania:inflation        -4.192e-03  2.127e-01  -0.020  0.98470
## continentSouth America:inflation   2.226e-01  1.427e-01   1.561  0.15307
## continentAsia:log_percapgdp        3.974e-01  1.605e+00   0.248  0.81006
## continentEurope:log_percapgdp     -1.609e+01  1.210e+01  -1.330  0.21636
## continentEurope/Asia:log_percapgdp        NA         NA      NA       NA
## continentNorth America:log_percapgdp -7.424e-01  1.780e+00  -0.417  0.68635
## continentOceania:log_percapgdp            NA         NA      NA       NA
## continentSouth America:log_percapgdp      NA         NA      NA       NA
## percapgdp:imports                 -8.483e-05  3.224e-05  -2.632  0.02729 *
## percapgdp:lpi                      3.422e-04  2.013e-04   1.700  0.12339
## percapgdp:inflation               -5.854e-06  1.765e-05  -0.332  0.74767
## percapgdp:log_percapgdp           -2.303e-04  2.218e-04  -1.038  0.32623
## imports:lpi                        1.113e+00  8.273e-01   1.345  0.21153
## imports:inflation                  1.354e-02  5.713e-02   0.237  0.81799
## imports:log_percapgdp              1.792e+00  6.749e-01   2.656  0.02622 *
## lpi:inflation                     -3.925e-02  5.097e-02  -0.770  0.46095
## lpi:log_percapgdp                 -2.234e+00  1.882e+00  -1.187  0.26548
## inflation:log_percapgdp           -2.454e-02  5.412e-02  -0.453  0.66093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3085 on 9 degrees of freedom
## Multiple R-squared:  0.9655, Adjusted R-squared:  0.793
## F-statistic: 5.598 on 45 and 9 DF,  p-value: 0.004624
```

Adding interaction terms increased our R^2 value to 0.793, which means the interaction terms help explain an additional 24% of the variation in banana prices. While this is a big improvement, the interactions between all of the continent indicator variables make our model overparameterized. The missing values indicate a lack of sufficient data, which makes sense given the model has 52 terms and our data set has 55 observations.

To reduce the size of our model, we'll go from a full second order model to a partial second order model, only including first order terms for each continent.

```
mod3 <- lm(price ~ (. -continent)^3 + continent, banana_modeling)

summary(mod3)
```

```
##
## Call:
## lm(formula = price ~ (. - continent)^3 + continent, data = banana_modeling)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.74511 -0.16078 -0.01012  0.15808  0.84124
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -2.469e+01  1.167e+01  -2.116   0.0454 *
## percapgdp                         -4.185e-03  3.566e-03  -1.173   0.2526
## imports                           -3.825e+00  1.311e+01  -0.292   0.7732
## lpi                                4.041e+00  6.095e+00   0.663   0.5140
## inflation                          1.075e+00  1.406e+00   0.765   0.4523
## log_percapgdp                      3.904e+00  1.665e+00   2.345   0.0280 *
## continentAsia                     -5.239e-01  2.909e-01  -1.801   0.0849 .
## continentEurope                   -5.742e-01  7.960e-01  -0.721   0.4780
## continentEurope/Asia              -1.047e+00  6.498e-01  -1.611   0.1208
## continentNorth America            -8.446e-02  2.769e-01  -0.305   0.7631
## continentOceania                   1.530e-01  4.000e-01   0.382   0.7057
## continentSouth America            -6.826e-01  3.841e-01  -1.777   0.0888 .
## percapgdp:imports                 -3.301e-04  8.615e-04  -0.383   0.7051
## percapgdp:lpi                      9.485e-04  6.929e-04   1.369   0.1842
## percapgdp:inflation                5.081e-04  3.694e-04   1.375   0.1823
## percapgdp:log_percapgdp            3.022e-04  3.222e-04   0.938   0.3581
## imports:lpi                       -1.361e+00  6.803e+00  -0.200   0.8432
## imports:inflation                  1.256e+00  7.474e-01   1.680   0.1065
## imports:log_percapgdp              5.407e-01  1.416e+00   0.382   0.7061
## lpi:inflation                      7.332e-01  1.109e+00   0.661   0.5152
## lpi:log_percapgdp                 -6.621e-01  7.815e-01  -0.847   0.4056
## inflation:log_percapgdp           -2.032e-01  1.708e-01  -1.190   0.2462
## percapgdp:imports:lpi             -2.709e-05  4.361e-05  -0.621   0.5406
## percapgdp:imports:inflation        1.209e-05  9.589e-06   1.260   0.2201
## percapgdp:imports:log_percapgdp    3.353e-05  8.258e-05   0.406   0.6885
## percapgdp:lpi:inflation           -2.672e-05  2.312e-05  -1.156   0.2596
## percapgdp:lpi:log_percapgdp       -6.792e-05  6.142e-05  -1.106   0.2802
## percapgdp:inflation:log_percapgdp -3.294e-05  3.570e-05  -0.923   0.3658
## imports:lpi:inflation              1.297e-01  7.859e-02   1.650   0.1125
## imports:lpi:log_percapgdp          1.525e-01  8.029e-01   0.190   0.8511
## imports:inflation:log_percapgdp   -1.986e-01  9.650e-02  -2.058   0.0511 .
## lpi:inflation:log_percapgdp       -9.205e-02  1.486e-01  -0.619   0.5417
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4231 on 23 degrees of freedom
## Multiple R-squared:  0.8342, Adjusted R-squared:  0.6107
## F-statistic: 3.733 on 31 and 23 DF,  p-value: 0.000866
```

This more concise model explains 0.182% less variation in price but the overall significance of our model has increased greatly (shown by the change in global p-value). This value compares our model performance with a baseline (intercept only) model. The lower the p-value, the comparatively better job our model does.

While our model does a moderate job of predicting prices, we can see that only three predictors are significant by any reasonable standard. To obtain a more concise model, we can try to pare down the model with backwards stepwise selection. By default, stepwise selection in R uses AIC which favors more complex models with higher explanatory power. Going backwards means we're removing the least important predictor (or predictor interaction) until the AIC of the model is maximized.

```
mod4 <- step(mod3, scope = price ~ ., trace = FALSE)

summary(mod4)
```

```
##
## Call:
## lm(formula = price ~ percapgdp + imports + lpi + inflation +
##     log_percapgdp + continent + percapgdp:imports + percapgdp:lpi +
##     percapgdp:inflation + percapgdp:log_percapgdp + imports:lpi +
##     imports:inflation + imports:log_percapgdp + lpi:inflation +
##     lpi:log_percapgdp + inflation:log_percapgdp + percapgdp:imports:lpi +
##     percapgdp:imports:inflation + percapgdp:lpi:inflation + percapgdp:inflation:log_percapgdp +
##     imports:lpi:inflation + imports:inflation:log_percapgdp,
##     data = banana_modeling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6728 -0.1874  0.0125  0.1661  0.7756
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       -1.750e+01  7.003e+00  -2.498  0.01886 *
## percapgdp                         -4.175e-04  6.929e-04  -0.602  0.55188
## imports                           -2.241e+00  1.526e+00  -1.468  0.15355
## lpi                                3.786e+00  4.107e+00   0.922  0.36480
## inflation                          1.659e+00  4.911e-01   3.379  0.00223 **
## log_percapgdp                      2.581e+00  8.447e-01   3.056  0.00501 **
## continentAsia                     -5.084e-01  2.560e-01  -1.986  0.05726 .
## continentEurope                   -3.374e-01  6.439e-01  -0.524  0.60463
## continentEurope/Asia              -1.043e+00  5.661e-01  -1.843  0.07630 .
## continentNorth America            -1.065e-01  2.530e-01  -0.421  0.67720
## continentOceania                   2.408e-01  3.590e-01   0.671  0.50809
## continentSouth America            -7.242e-01  3.472e-01  -2.086  0.04655 *
## percapgdp:imports                  4.285e-05  2.892e-05   1.482  0.14995
## percapgdp:lpi                      1.844e-04  6.746e-05   2.734  0.01090 *
## percapgdp:inflation                2.275e-04  8.320e-05   2.734  0.01091 *
## percapgdp:log_percapgdp           -2.939e-05  6.803e-05  -0.432  0.66914
## imports:lpi                       -6.723e-02  4.260e-01  -0.158  0.87580
## imports:inflation                  7.590e-01  3.455e-01   2.197  0.03681 *
## imports:log_percapgdp              2.820e-01  1.720e-01   1.639  0.11277
## lpi:inflation                      3.558e-02  6.448e-02   0.552  0.58559
## lpi:log_percapgdp                 -5.514e-01  4.943e-01  -1.116  0.27445
## inflation:log_percapgdp           -2.533e-01  7.817e-02  -3.240  0.00316 **
## percapgdp:imports:lpi             -1.480e-05  8.830e-06  -1.677  0.10518
## percapgdp:imports:inflation        6.635e-06  5.027e-06   1.320  0.19801
## percapgdp:lpi:inflation           -3.111e-05  1.473e-05  -2.111  0.04413 *
## percapgdp:inflation:log_percapgdp -7.602e-06  7.532e-06  -1.009  0.32183
## imports:lpi:inflation              1.153e-01  7.142e-02   1.614  0.11811
## imports:inflation:log_percapgdp   -1.322e-01  4.660e-02  -2.836  0.00855 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4018 on 27 degrees of freedom
## Multiple R-squared:  0.8244, Adjusted R-squared:  0.6489
## F-statistic: 4.696 on 27 and 27 DF,  p-value: 7.015e-05
```

With the backwards stepwise selection, our R^2 value increases to 0.649 and we've removed 4 predictors. Instead of backwards selection, we can also try a bidirectional stepwise selection where the model can drop or add predictors at each step, starting from `mod3`. This should produce the simplest, most powerful model.

```
mod5 <- step(mod3, scope = price ~ ., direction = "both", trace = FALSE)

summary(mod5)
```

```
##
## Call:
## lm(formula = price ~ percapgdp + imports + lpi + inflation +
##     log_percapgdp + continent + percapgdp:imports + percapgdp:lpi +
##     percapgdp:inflation + percapgdp:log_percapgdp + imports:lpi +
##     imports:inflation + imports:log_percapgdp + lpi:inflation +
##     lpi:log_percapgdp + inflation:log_percapgdp + percapgdp:imports:lpi +
##     percapgdp:imports:inflation + percapgdp:lpi:inflation + percapgdp:inflation:log_percapgdp +
##     imports:lpi:inflation + imports:inflation:log_percapgdp,
##     data = banana_modeling)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6728 -0.1874  0.0125  0.1661  0.7756
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -1.750e+01  7.003e+00  -2.498  0.01886 *
## percapgdp                       -4.175e-04  6.929e-04  -0.602  0.55188
## imports                         -2.241e+00  1.526e+00  -1.468  0.15355
## lpi                              3.786e+00  4.107e+00   0.922  0.36480
## inflation                        1.659e+00  4.911e-01   3.379  0.00223 **
## log_percapgdp                    2.581e+00  8.447e-01   3.056  0.00501 **
## continentAsia                   -5.084e-01  2.560e-01  -1.986  0.05726 .
## continentEurope                 -3.374e-01  6.439e-01  -0.524  0.60463
## continentEurope/Asia            -1.043e+00  5.661e-01  -1.843  0.07630 .
## continentNorth America          -1.065e-01  2.530e-01  -0.421  0.67720
## continentOceania                 2.408e-01  3.590e-01   0.671  0.50809
## continentSouth America          -7.242e-01  3.472e-01  -2.086  0.04655 *
## percapgdp:imports                4.285e-05  2.892e-05   1.482  0.14995
## percapgdp:lpi                    1.844e-04  6.746e-05   2.734  0.01090 *
## percapgdp:inflation              2.275e-04  8.320e-05   2.734  0.01091 *
## percapgdp:log_percapgdp         -2.939e-05  6.803e-05  -0.432  0.66914
## imports:lpi                     -6.723e-02  4.260e-01  -0.158  0.87580
## imports:inflation                7.590e-01  3.455e-01   2.197  0.03681 *
## imports:log_percapgdp            2.820e-01  1.720e-01   1.639  0.11277
## lpi:inflation                    3.558e-02  6.448e-02   0.552  0.58559
## lpi:log_percapgdp               -5.514e-01  4.943e-01  -1.116  0.27445
## inflation:log_percapgdp         -2.533e-01  7.817e-02  -3.240  0.00316 **
## percapgdp:imports:lpi           -1.480e-05  8.830e-06  -1.677  0.10518
## percapgdp:imports:inflation      6.635e-06  5.027e-06   1.320  0.19801
## percapgdp:lpi:inflation         -3.111e-05  1.473e-05  -2.111  0.04413 *
## percapgdp:inflation:log_percapgdp -7.602e-06  7.532e-06  -1.009  0.32183
## imports:lpi:inflation            1.153e-01  7.142e-02   1.614  0.11811
## imports:inflation:log_percapgdp -1.322e-01  4.660e-02  -2.836  0.00855 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4018 on 27 degrees of freedom
## Multiple R-squared:  0.8244, Adjusted R-squared:  0.6489
## F-statistic: 4.696 on 27 and 27 DF,  p-value: 7.015e-05
```

Using bidirectional stepwise selection, the same predictor set was chosen. It seems like this is the optimal model for our purposes. During tuning we've substantially increased the proportion of variation our model explains while keeping it as concise as possible.

Cross Validation

But wait! If you're familiar with regression you know that the judging a model's performance by it's predictions of training data is not the best practice because it can lead to overfitting. Instead, we should test the model's accuracy on a country it hasn't seen before.

We're going to use leave out one cross validation (LOOCV), which excludes one country from the model fitting process and then tests the model's accuracy on that country. The process is then repeated for each country in the data set, averaging the test error until we have a representative error metric.

I'm using the built-in LOOCV functionality included in the `caret` package to perform cross validation on model 5. I choose to perform LOOCV over K-fold because we don't have many observations in the data set and I saw large variations in model coefficients split-to-split. LOOCV is more stable and makes sense for our purposes, at the expense of being more computationally taxing.

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
```

```
train_ctrl <- trainControl(method = "cv", number = 10)

ideal_formula <- mod5$call[[2]]

full_mod <- train(ideal_formula, data = banana_modeling, method = "lm", trControl = train_ctrl)

print(full_mod)
```

```
## Linear Regression
##
## 55 samples
##  6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 50, 50, 48, 49, 51, 49, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   2.265475  0.4493011  1.414227
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

We can see that the cross-validated R^2 is 0.2 lower than our training R^2. We expect to see a hit in accuracy because the model is predicting a value not seen during its training. CV test error is crucial in evaluating a model's real-world performance; it doesn't make sense to predict a value you've already collected.

Analyzing Model Performance

Now that we have a tuned model, I'm curious to see how well it predicts banana prices. According to the LOOP, we would expect to see roughly the same price everywhere. If our model does a good job predicting prices, we have some evidence against the LOOP's credibility.

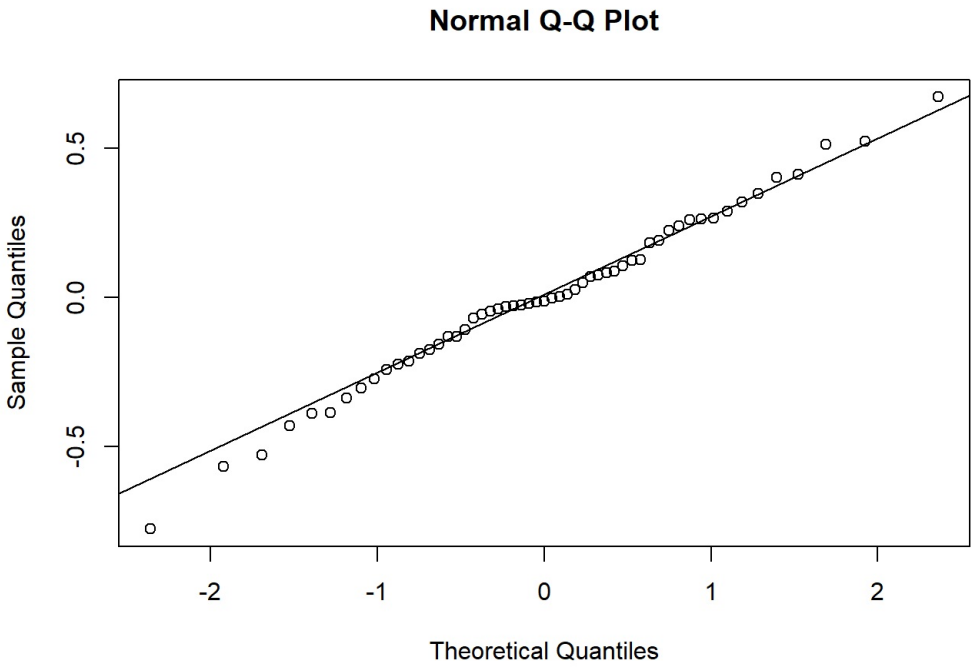
First, let's check the largest discrepancies between the model and the real data:

```
banana_preds <- predict(full_mod, newdata = banana_modeling, type = "raw")

banana_dev <- bananas |>
              mutate(prediction = banana_preds,
                     deviation = prediction - price,
                     abs_dev = abs(deviation)) |>
              arrange(desc(abs_dev)) |>
              select(country, price, prediction, deviation, abs_dev)

head(banana_dev, n = 10)
```

```
##                country      price prediction  deviation   abs_dev
## 1               Panama 2.20000000  1.4244067 -0.7755933 0.7755933
## 2             Tanzania 0.09166667  0.7645001  0.6728334 0.6728334
## 3          Ivory Coast 1.50000000  0.9326684 -0.5673316 0.5673316
## 4               Zambia 1.66666667  1.1403210 -0.5263457 0.5263457
## 5  United Arab Emirates 0.48461748  1.0088869  0.5242694 0.5242694
## 6              Nigeria 0.27642276  0.7908525  0.5144297 0.5144297
## 7                Ghana 1.16000000  0.7309112 -0.4290888 0.4290888
## 8           Philippines 0.02653094  0.4396261  0.4130951 0.4130951
## 9               Jamaica 0.81911921  1.2219636  0.4028444 0.4028444
## 10           Bangladesh 0.77272727  0.3832915 -0.3894358 0.3894358
```

A cursory glance at the deviation column shows that our model seems to over and under estimate prices at roughly equal frequencies. In general, we'd expect the residuals to lie equally above and below the true values. We can check this using a residual plot.

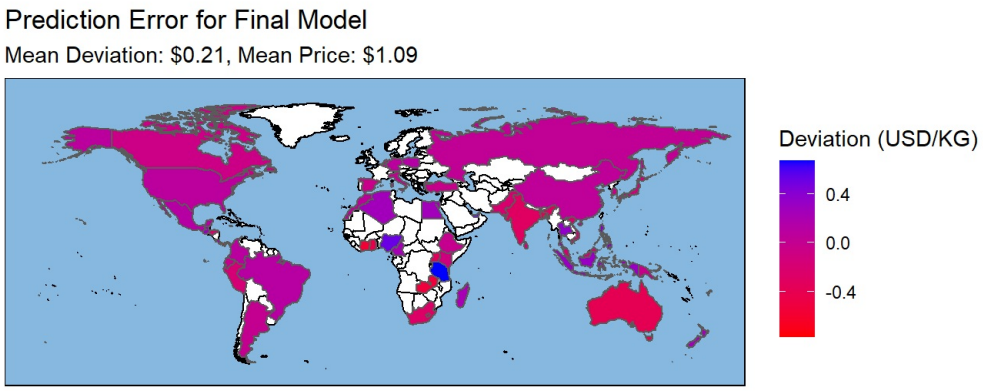
```
qqnorm(banana_dev$deviation)
qqline(banana_dev$deviation)
```



When examining a QQ-plot, we'd like the residuals to lie perfectly on the line which would indicate our data has a perfectly normal distribution. While it seems this is generally true, the sinusoidal waviness of the tails of the residuals makes me think there might be a higher order trend at play. In a future addition, I'm going to test a ridge or LASSO model which would allow the model to take different curvature depending on the data's distribution.

We can also examine the deviation visually using our beloved world map. I'm forgoing the banana color scheme to improve interpretability; we'll represent negative deviations (over estimations) with red and positive deviations (under estimations) with blue.

```
mean_dev <- round(mean(abs(banana_dev$deviation)), 2)
mean_price <- round(mean(banana_dev$price), 2)

subt <- paste0("Mean Deviation: $", mean_dev, ", Mean Price: $", mean_price)

dev_plot <- ggplot() +
            geom_map(data = world_coords, map = world_coords,
                     aes(group = group, map_id = region),
                         fill = "white", color = "black", linewidth = 0.5) +
              geom_map(data = banana_dev, map = world_coords,
                       aes(fill = deviation, map_id = country),
                           color="#5b5b5b", linewidth = 0.5) +
              coord_map("rectangular", lat0=0, xlim = c(-180,180), ylim = c(-60, 90)) +
              scale_fill_continuous(low="red", high="blue", guide = "colorbar") +
              labs(fill = "Deviation (USD/KG)") +
              ggtitle("Prediction Error for Final Model",
                      subtitle = subt) +
              ylab("") + xlab("") +
              theme_minimal() +
              theme(panel.background = element_rect(fill = '#87B8DF', color = 'black'))

dev_plot
```




```
#dev.copy(device = png, filename = "deviation_plot.png", width = 1500, height = 750)
#dev.off()
```

Broadly, we see that the model tends to underestimate the cost for North America, Europe, and Oceania, while it's more accurate for South America and South Asia. As we noticed in the EDA phase, Africa has a wide variety in banana prices and the model struggles to predict prices accurately.

Part 3: Conclusion:

With our final model accounting for logistics performance, income differences, and other factors, we can explain half of all variation in banana pricing. This is a somewhat unsatisfying result, as the LOOP suggests we should be able to explain almost all of the variation in pricing if we account for shipping and transaction costs. Our model's ability to explain some of the variation suggests that the LOOP applies in some capacity, but doesn't hold absolutely.

It is known that the LOOP does not hold in practice, often due to over/undervalued currencies or government intervention in free trade. This seems consistent with our result, where we can partially (but not totally) explain fluctuations in pricing.

Part 4: Next Steps

To further bolster our analysis, I plan to consider the following:

- Ridge/LASSO/PSR for curve smoothing in the regression model
- Adding predictors for tax and tariff rates
- Experimenting with KNN and regression trees (would require more data)