Probability with R - Exercise 12.2

Lucas Liona

Table of contents

Problem

Duncan (1994) looked at drug law enforcement expenditures and drug-induced deaths. Table 12.2.2 gives figures from 1981 to 1991 on the U.S. DEA (Drug Enforcement Agency) budget and the numbers of drug-induced deaths in the United States.

(a) Plot deaths versus DEA budget. Do you think the budget causes deaths? Why not? Plot budget versus deaths. What do you think now?

(b) The variables deaths and budget are affected by a third variable—year.

  (i) Plot budget versus year. Do you think that a straight line would adequately fit this scatter plot?
  (ii) What other trend might fit?

(c) Plot deaths versus year. What do you conclude from the three plots?

Data Entry and Preparation

First we have to input the data from the table and create a data frame:

```
# Create vectors for the data
year <- c(1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991)
budget <- c(216, 239, 255, 292, 344, 372, 486, 493, 543, 558, 692)
deaths <- c(7106, 7310, 7492, 7892, 8663, 9976, 9796, 10917, 10710, 9463, 10388)

# Create a data frame
dea_data <- data.frame(year, budget, deaths)

# Display the data
knitr::kable(dea_data,
             col.names = c("Year", "Budget ($ millions)", "Deaths"),
             caption = "DEA Budget and Drug Deaths (1981-1991)")
```

Table 1: DEA Budget and Drug Deaths (1981-1991)

Year	Budget ($ millions)	Deaths
1981	216	7106
1982	239	7310
1983	255	7492
1984	292	7892
1985	344	8663
1986	372	9976
1987	486	9796
1988	493	10917
1989	543	10710
1990	558	9463
1991	692	10388

Basic Summary Statistics

Before visualizing the data we can examine some basic statistics/properties:

```
# Basic summary statistics
summary(dea_data)
```

```
      year          budget          deaths
```

```
 Min.   :1981   Min.   :216.0   Min.   : 7106
 1st Qu.:1984   1st Qu.:273.5   1st Qu.: 7692
 Median :1986   Median :372.0   Median : 9463
 Mean   :1986   Mean   :408.2   Mean   : 9065
 3rd Qu.:1988   3rd Qu.:518.0   3rd Qu.:10182
 Max.   :1991   Max.   :692.0   Max.   :10917
```

```r
# Correlation between variables
cor_matrix <- cor(dea_data)
knitr::kable(cor_matrix, digits = 3,
             caption = "Correlation Matrix for DEA Data")
```

Table 2: Correlation Matrix for DEA Data

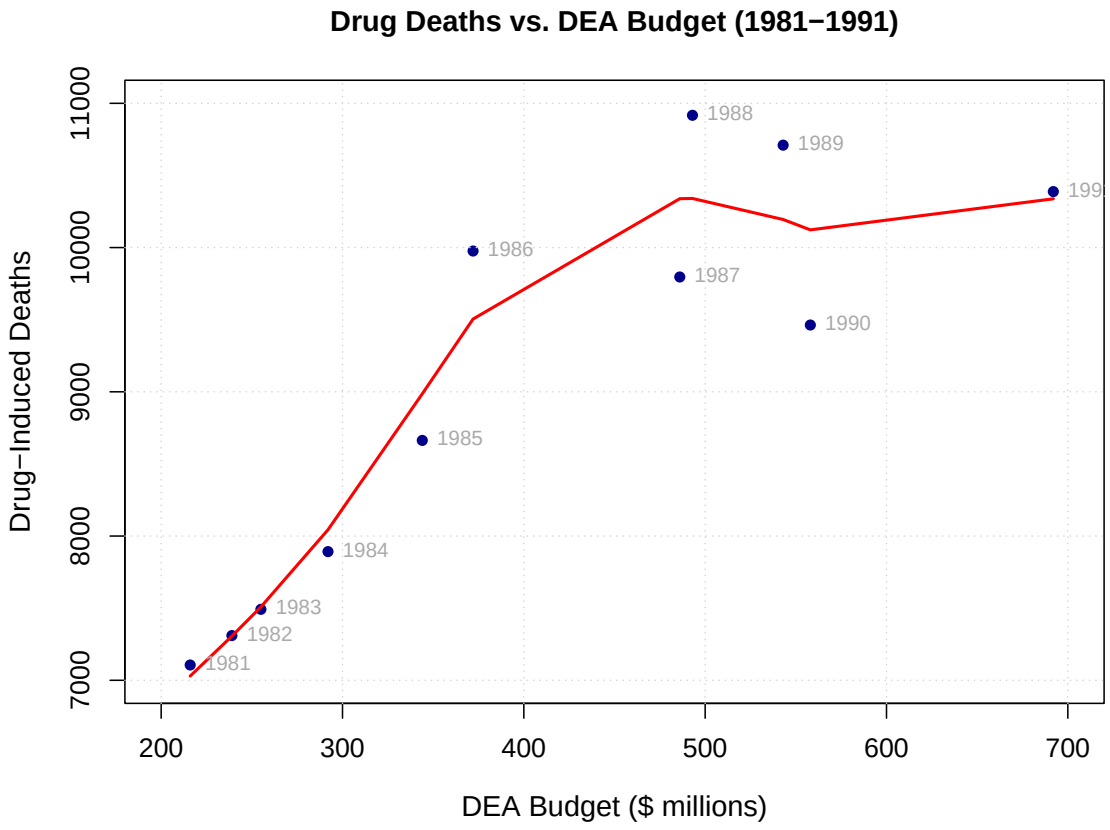
	year	budget	deaths
year	1.000	0.981	0.885
budget	0.981	1.000	0.862
deaths	0.885	0.862	1.000

Part (a): Deaths versus DEA Budget

This is a scatter plot of deaths versus DEA budget:

```r
# Set plot parameters for better readability
par(mar = c(5, 5, 4, 2), cex.lab = 1.2, cex.axis = 1.1)

# Plot deaths versus budget
plot(dea_data$budget, dea_data$deaths,
     main = "Drug Deaths vs. DEA Budget (1981-1991)",
     xlab = "DEA Budget ($ millions)",
     ylab = "Drug-Induced Deaths",
     pch = 16,
     col = "darkblue",
     xlim = c(200, 700),
     ylim = c(7000, 11000))

# Add grid lines for better readability
grid(lty = "dotted", col = "lightgray")

# Add a smooth trend line
```

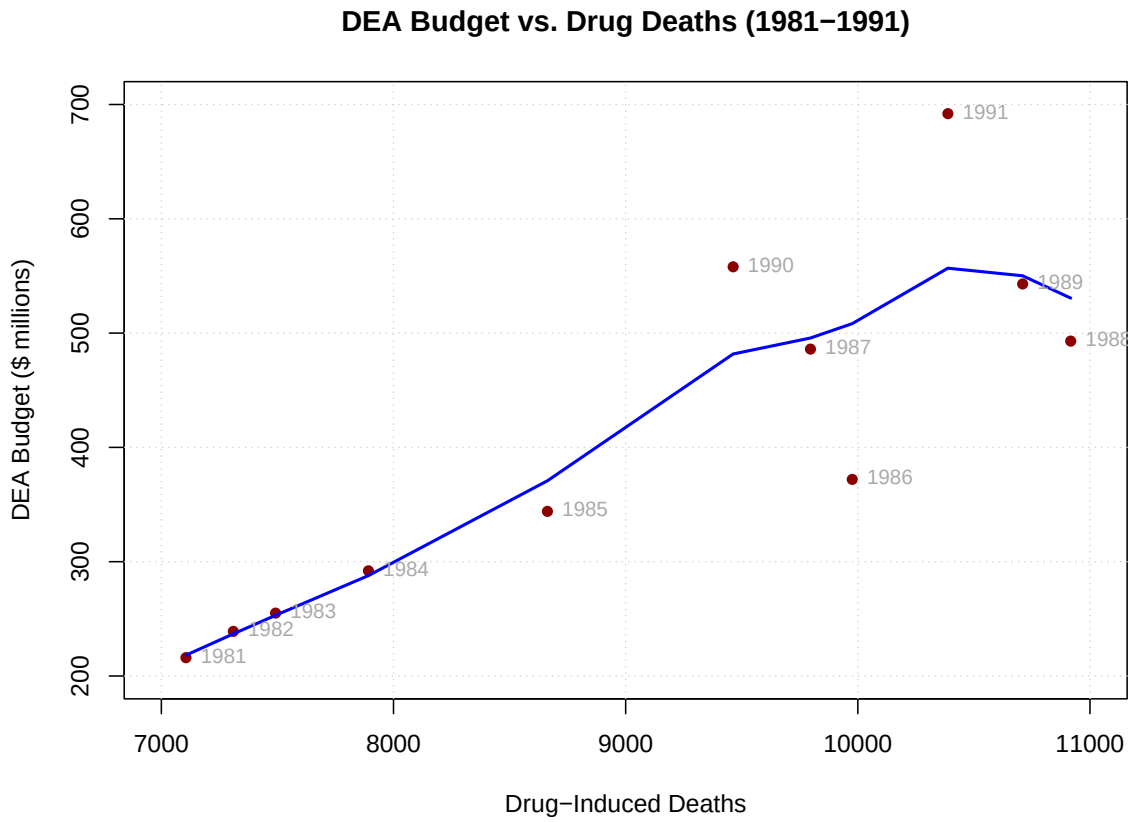
```r
lines(smooth.spline(dea_data$budget, dea_data$deaths, df = 5),
      col = "red", lwd = 2)

# Add text labels showing the year for each point
text(dea_data$budget, dea_data$deaths,
     labels = dea_data$year,
     pos = 4,
     cex = 0.8,
     col = "darkgray")
```



Now let's plot budget versus deaths (reverse axes) to see a different persepctive about their relationship:

```r
# Plot budget versus deaths
plot(dea_data$deaths, dea_data$budget,
     main = "DEA Budget vs. Drug Deaths (1981-1991)",
     xlab = "Drug-Induced Deaths",
```

```
     ylab = "DEA Budget ($ millions)",
     pch = 16,
     col = "darkred",
     ylim = c(200, 700),
     xlim = c(7000, 11000))

# Add grid lines
grid(lty = "dotted", col = "lightgray")

# Add a smooth trend line
lines(smooth.spline(dea_data$deaths, dea_data$budget, df = 5),
      col = "blue", lwd = 2)

# Add text labels showing the year for each point
text(dea_data$deaths, dea_data$budget,
     labels = dea_data$year,
     pos = 4,
     cex = 0.8,
     col = "darkgray")
```



Interpretation:

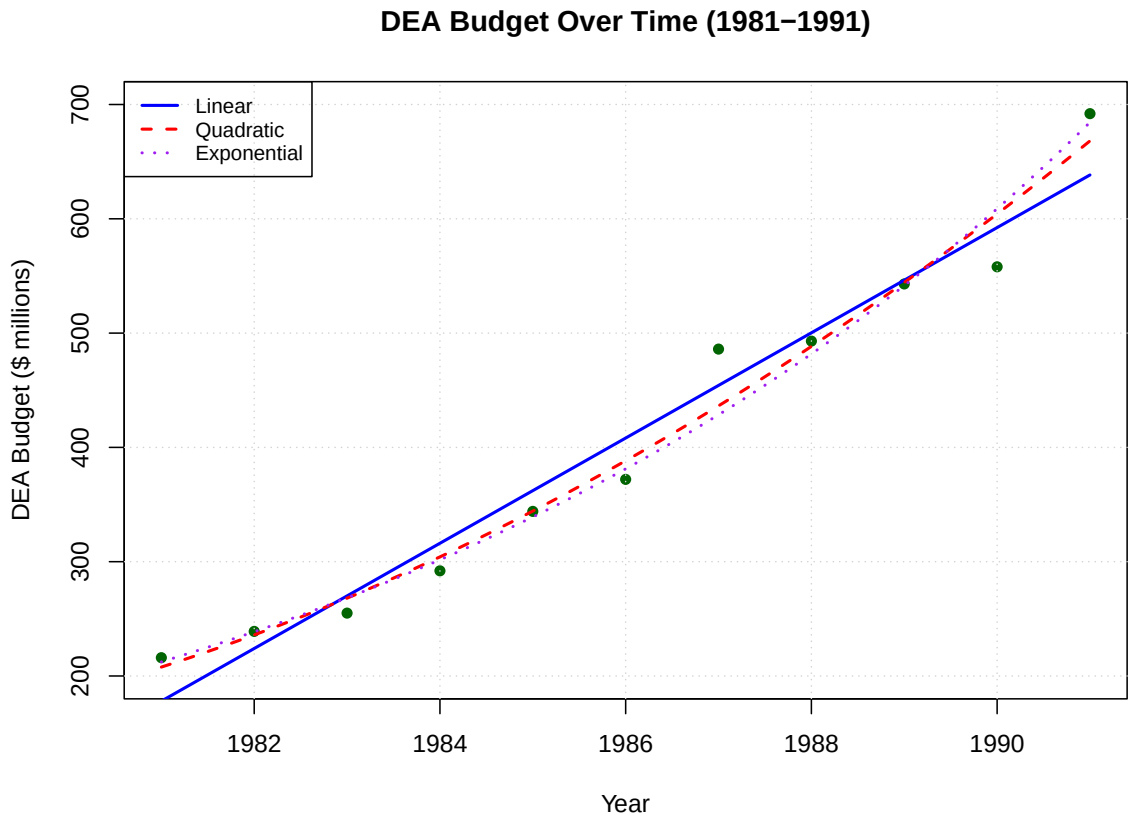
The plot of deaths versus DEA budget shows a positive association between the two variables. But this does not necessarily mean that the budget causes deaths. Correlation does not imply causation. Both variables might be increasing over time independently or they could be influenced by other factors. (See Lurking Variable and Conclusions for more on this)

When we plot budget versus deaths, we see the same relationship from a different perspective. The budget did not cause the deaths as a statistical relationship does not necessarily imply a causal relationship. If drug problems are getting worse over time (reflected in increased numbers of deaths) and budgets are continually increased to try to combat the problem, this would be a plausible explanation and a pattern/trend similar to this could be expected.

Part (b): Budget versus Year

Now we can examine how the DEA budget has changed over time (our lurking variable):

```r
# Plot budget versus year
plot(dea_data$year, dea_data$budget,
     main = "DEA Budget Over Time (1981-1991)",
     xlab = "Year",
     ylab = "DEA Budget ($ millions)",
     pch = 16,
     col = "darkgreen",
     ylim = c(200, 700),
     xlim = c(1981, 1991))

# Add grid lines
grid(lty = "dotted", col = "lightgray")

# Fit linear, quadratic, and exponential models
linear_model <- lm(budget ~ year, data = dea_data)
quadratic_model <- lm(budget ~ year + I(year^2), data = dea_data)
exp_model <- lm(log(budget) ~ year, data = dea_data)

# Add trend lines for different models
years_seq <- seq(1981, 1991, 0.1)
lines(dea_data$year, fitted(linear_model), col = "blue", lwd = 2)
lines(years_seq, predict(quadratic_model, newdata = data.frame(year = years_seq)),
      col = "red", lwd = 2, lty = 2)
lines(years_seq, exp(predict(exp_model, newdata = data.frame(year = years_seq))),
      col = "purple", lwd = 2, lty = 3)

# Add a legend
legend("topleft",
       legend = c("Linear", "Quadratic", "Exponential"),
       col = c("blue", "red", "purple"),
       lty = c(1, 2, 3),
       lwd = 2,
       cex = 0.8)
```



Let's compare the models statistically:

```
# Print model summaries
cat("Linear Model (Budget ~ Year):\n")
```

```
Linear Model (Budget ~ Year):
```

```
summary(linear_model)
```

```

Call:
lm(formula = budget ~ year, data = dea_data)

Residuals:
    Min      1Q  Median      3Q     Max
-36.182 -21.127  -7.255  23.373  53.636
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -91020.036   6068.484  -15.00 1.13e-07 ***
year            46.036      3.056   15.07 1.09e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.05 on 9 degrees of freedom
Multiple R-squared:  0.9619,    Adjusted R-squared:  0.9576
F-statistic:   227 on 1 and 9 DF,  p-value: 1.086e-07
```

```
cat("\nQuadratic Model (Budget ~ Year + Year²):\n")
```

```
Quadratic Model (Budget ~ Year + Year²):
```

```
summary(quadratic_model)
```

```
Call:
lm(formula = budget ~ year + I(year^2), data = dea_data)

Residuals:
    Min      1Q  Median      3Q     Max
-46.243 -12.640  -0.271   6.436  49.656

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.742e+06  3.644e+06   2.125   0.0664 .
year        -7.842e+03  3.670e+03  -2.137   0.0651 .
I(year^2)    1.986e+00  9.239e-01   2.150   0.0638 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.06 on 8 degrees of freedom
Multiple R-squared:  0.9758,    Adjusted R-squared:  0.9698
F-statistic: 161.5 on 2 and 8 DF,  p-value: 3.416e-07
```

```
cat("\nExponential Model (log(Budget) ~ Year):\n")
```

```
Exponential Model (log(Budget) ~ Year):
```

```
summary(exp_model)
```

```
Call:
lm(formula = log(budget) ~ year, data = dea_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.087147 -0.028512  0.002638  0.015692  0.125805

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.265e+02  1.072e+01  -21.13 5.59e-09 ***
year         1.170e-01  5.397e-03   21.68 4.45e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05661 on 9 degrees of freedom
Multiple R-squared:  0.9812,    Adjusted R-squared:  0.9791
F-statistic: 470.2 on 1 and 9 DF,  p-value: 4.447e-09
```

```
# Compare models using AIC
AIC(linear_model, quadratic_model, exp_model)
```

```
                df       AIC
linear_model     3 111.28822
quadratic_model  4 108.27348
exp_model        3 -28.16595
```

Interpretation:

The plot of budget versus year shows a clear increasing trend over time. A straight line does not adequately fit this scatter plot - there appears to be a non-linear relationship where the budget increases more rapidly in later years.

Other trends that might fit better are

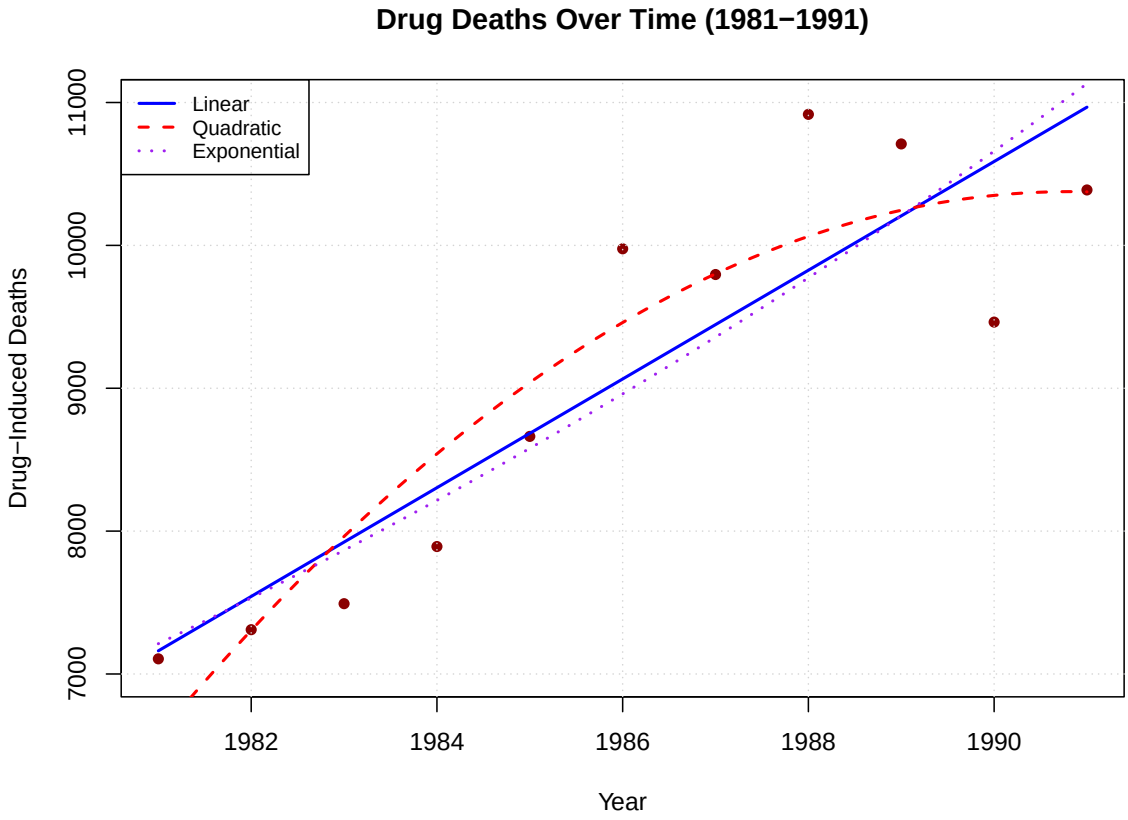
1. Exponential growth (where the budget grows by a percentage each year)
2. Quadratic growth (where the rate of increase itself increases over time)

Based on the AIC values and though visual observation, we can assume these are the fitting trends.

Part (c): Deaths versus Year

Finally we examine how drug-induced deaths have changed over time:

```
# Plot deaths versus year
plot(dea_data$year, dea_data$deaths,
     main = "Drug Deaths Over Time (1981-1991)",
     xlab = "Year",
     ylab = "Drug-Induced Deaths",
     pch = 16,
     col = "darkred",
     ylim = c(7000, 11000),
     xlim = c(1981, 1991))

# Add grid lines
grid(lty = "dotted", col = "lightgray")

# Fit linear, quadratic, and exponential models for deaths vs year
linear_model_deaths <- lm(deaths ~ year, data = dea_data)
quadratic_model_deaths <- lm(deaths ~ year + I(year^2), data = dea_data)
exp_model_deaths <- lm(log(deaths) ~ year, data = dea_data)

# Add trend lines for different models
lines(dea_data$year, fitted(linear_model_deaths), col = "blue", lwd = 2)
lines(years_seq, predict(quadratic_model_deaths, newdata = data.frame(year = years_seq)),
      col = "red", lwd = 2, lty = 2)
lines(years_seq, exp(predict(exp_model_deaths, newdata = data.frame(year = years_seq))),
      col = "purple", lwd = 2, lty = 3)

# Add a legend
legend("topleft",
       legend = c("Linear", "Quadratic", "Exponential"),
       col = c("blue", "red", "purple"),
       lty = c(1, 2, 3),
       lwd = 2,
       cex = 0.8)
```



Comparing the models for deaths over time:

```
# Print model summaries
cat("Linear Model (Deaths ~ Year):\n")
```

```
Linear Model (Deaths ~ Year):
```

```
summary(linear_model_deaths)
```

```
Call:
lm(formula = deaths ~ year, data = dea_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1123.96  -421.48   -56.14   427.11  1091.11
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -746680.40  132545.13  -5.633 0.000320 ***
year            380.54      66.74   5.702 0.000294 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 700 on 9 degrees of freedom
Multiple R-squared:  0.7832,    Adjusted R-squared:  0.7591
F-statistic: 32.51 on 1 and 9 DF,  p-value: 0.0002936
```

```
cat("\nQuadratic Model (Deaths ~ Year + Year²):\n")
```

```
Quadratic Model (Deaths ~ Year + Year²):
```

```
summary(quadratic_model_deaths)
```

```
Call:
lm(formula = deaths ~ year + I(year^2), data = dea_data)

Residuals:
    Min      1Q  Median      3Q     Max
-886.91 -423.79    4.38  490.08  854.05

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.566e+08  8.342e+07  -1.877   0.0973 .
year         1.573e+05  8.401e+04   1.873   0.0980 .
I(year^2)   -3.951e+01  2.115e+01  -1.868   0.0987 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 619.5 on 8 degrees of freedom
Multiple R-squared:  0.849, Adjusted R-squared:  0.8113
F-statistic:  22.5 on 2 and 8 DF,  p-value: 0.0005193
```

```r
cat("\nExponential Model (log(Deaths) ~ Year):\n")
```

```
Exponential Model (log(Deaths) ~ Year):
```

```r
summary(exp_model_deaths)
```

```
Call:
lm(formula = log(deaths) ~ year, data = dea_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.11903 -0.04442 -0.01480  0.04697  0.11072

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -77.110390  14.442205  -5.339 0.000469 ***
year          0.043409   0.007272   5.969 0.000210 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07627 on 9 degrees of freedom
Multiple R-squared:  0.7984,    Adjusted R-squared:  0.776
F-statistic: 35.63 on 1 and 9 DF,  p-value: 0.0002103
```

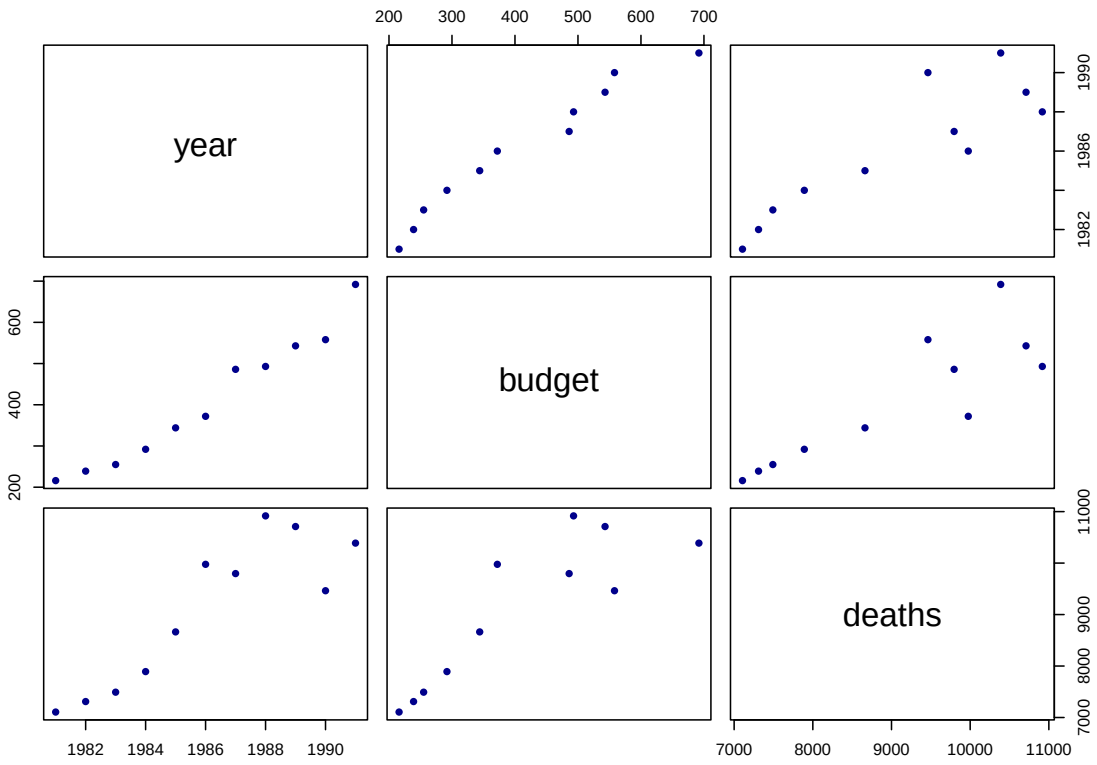
```r
# Compare models using AIC
AIC(linear_model_deaths, quadratic_model_deaths, exp_model_deaths)
```

```
                       df       AIC
linear_model_deaths     3 179.13214
quadratic_model_deaths  4 177.14999
exp_model_deaths        3 -21.60738
```

Conclusions

Based on the three plots, we can draw several conclusions:

```r
# Create a scatter plot matrix for a comprehensive overview
pairs(dea_data,
      main = "Scatter Plot Matrix: DEA Budget, Deaths, and Year",
      pch = 16,
      col = "darkblue")
```



1. **Deaths vs. Budget**: There is a positive correlation between DEA budget and drug deaths (r = 0.86). However, this correlation likely does not indicate causation. The budget did not cause the deaths.
2. **Budget vs. Year**: The DEA budget increased over time in a non-linear fashion. Both exponential and quadratic models fit better than a linear model, suggesting accelerating budget growth.
3. **Deaths vs. Year**: Drug-induced deaths also increased over time, but with notable fluctuations, particularly in 1989-1991.

4. **Overall Conclusion**: There does not seem to be any relationship between budget and deaths other than that they both tend to increase with time. Both variables increase independently over time, which explains their apparent correlation. The data does not provide evidence that increasing the DEA budget reduces drug deaths - in fact, despite the increasing budget, deaths generally continued to rise over this period.
5. **Potential Implications**: This raises questions about the effectiveness of increased DEA funding in reducing drug-related mortality during this period. However, we should be cautious about drawing policy conclusions from correlation analysis alone. There may be confounding variables not captured in this dataset, or there may be time lags between policy changes and their effects.

Lurking Variable

Yes, there is a lurking variable, and in this case it's time. As you can see in the scatter-plot matrix, while deaths and budget do increase in an almost linear way, implying a correlation, there is a lurking variable time that is responsible for this correlation, and we know correlation =/= causation. You can also see in the scatter-plot matrix that budget linearly increases with time, and so do deaths.