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
```

```
# 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

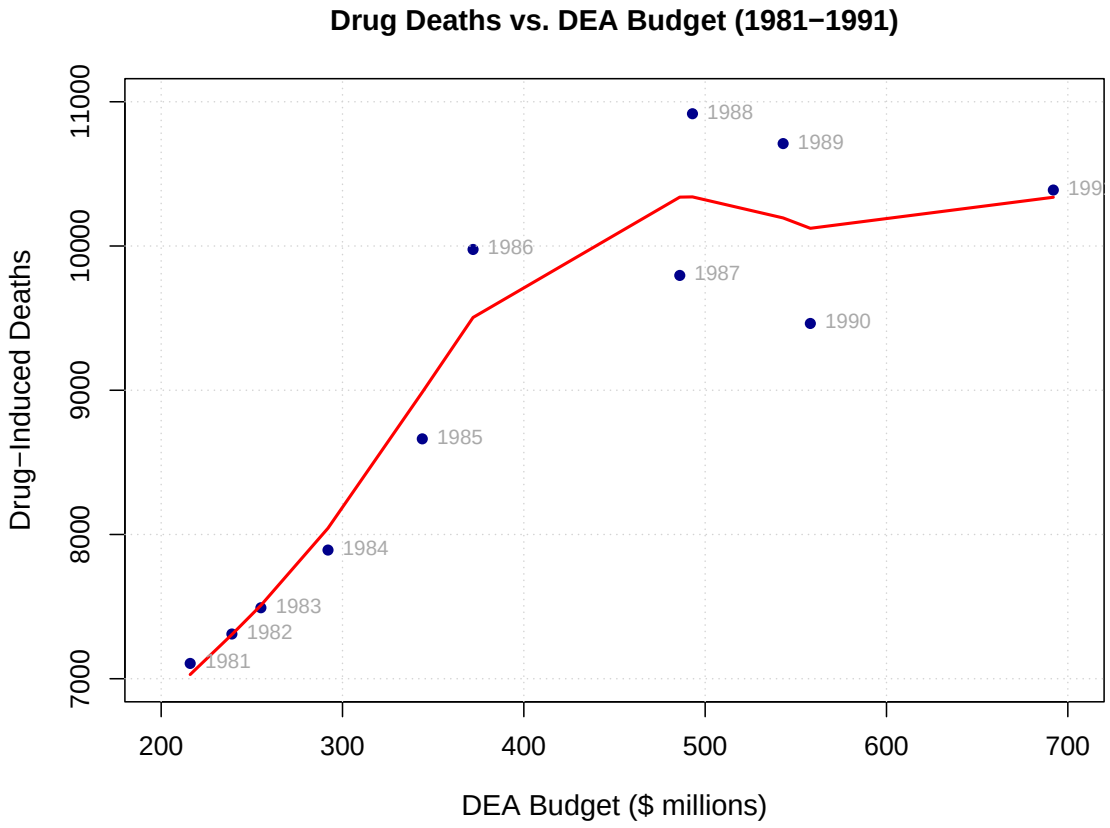
	year	budget	deaths
year	1.000	0.981	0.885
budget	0.981	1.000	0.862
deaths	0.885	0.862	1.000

Part (a): Deaths versus DEA Budget

This is a scatter plot of deaths versus DEA budget:

```
# 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,
```

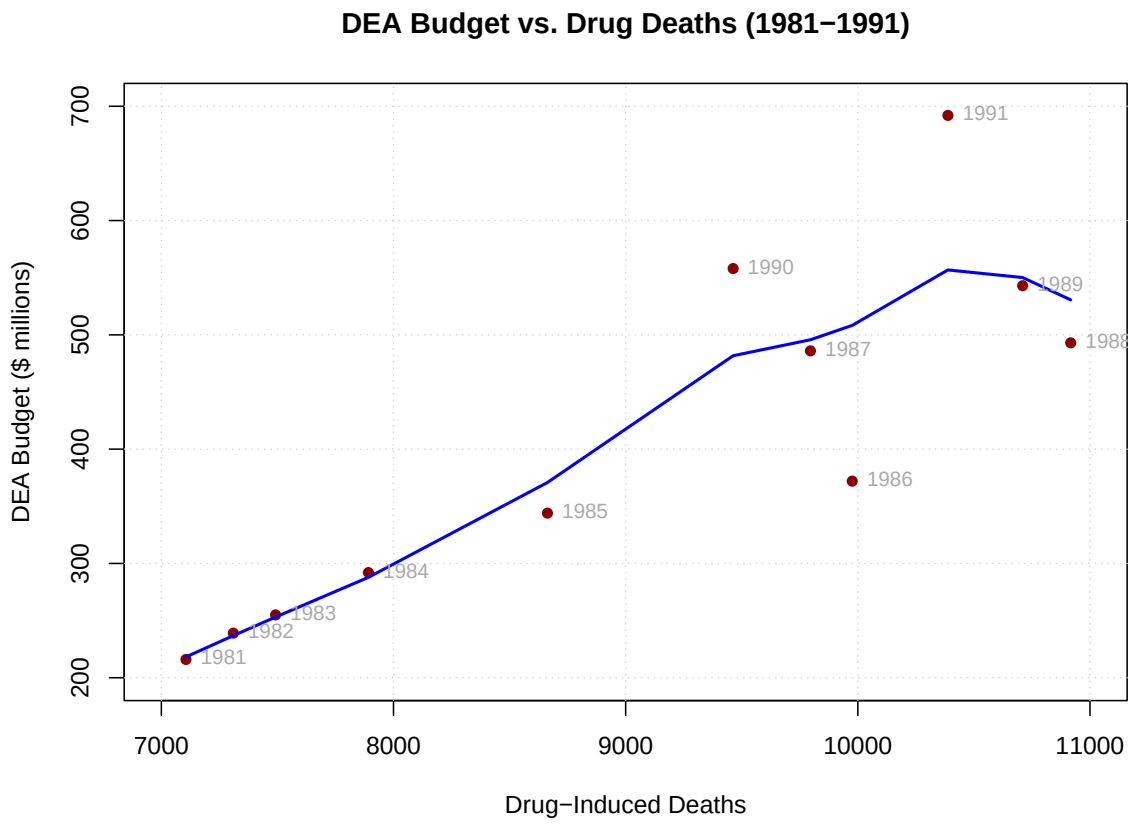
```
     col = "darkblue",
     xlim = c(200, 700),
     ylim = c(7000, 11000))

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

# Add a smooth trend line
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 perspective about their relationship:

```
# 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
```

```r
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)

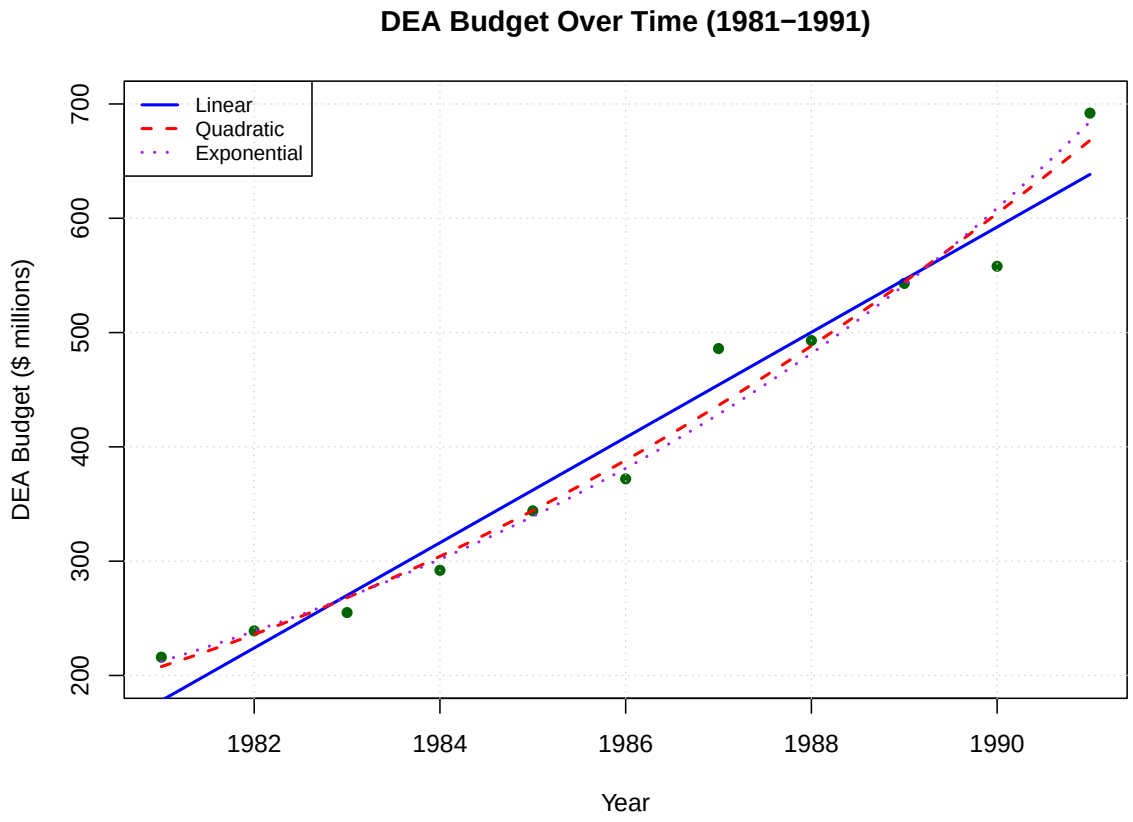
When we plot budget versus deaths, we see the same relationship from a different perspective. The budget did not cause the deaths as a statistical relationship does not necessarily imply a causal relationship. If drug problems are getting worse over time (reflected in increased numbers of deaths) and 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):

```
# 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:

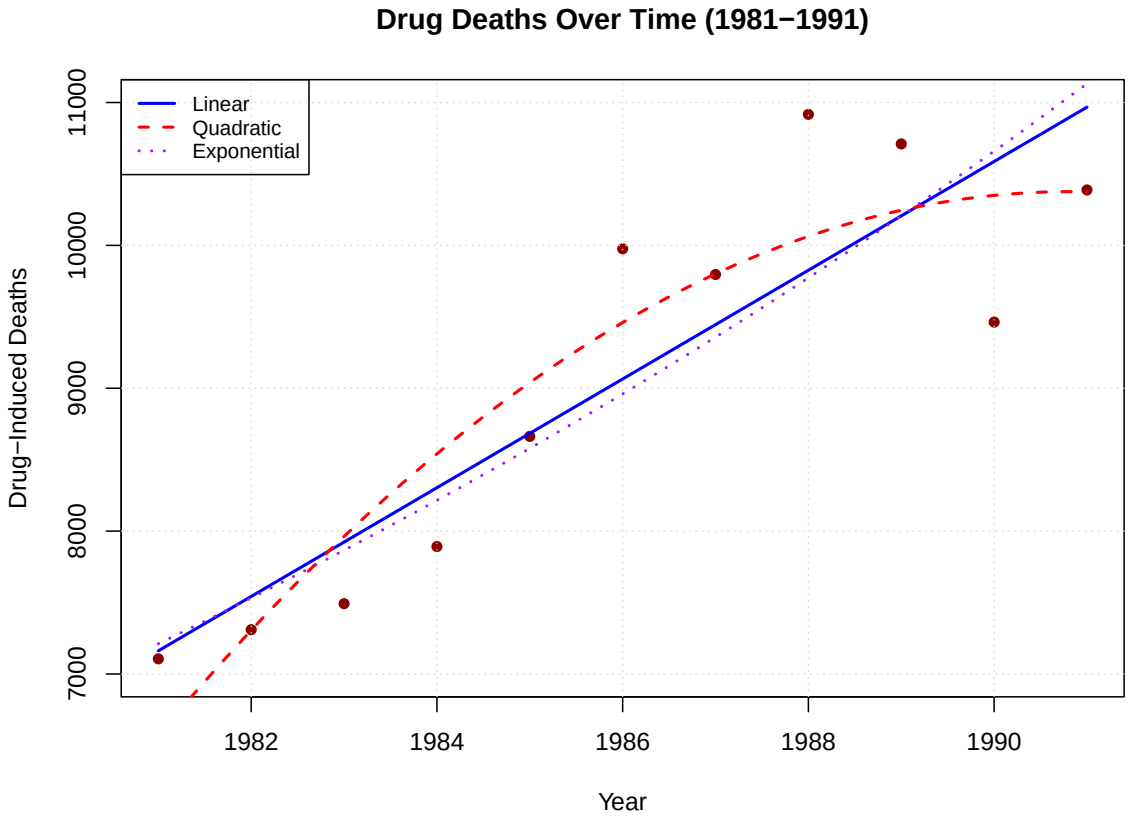
1. Exponential growth (where the budget grows by a percentage each year)
2. Quadratic growth (where the rate of increase itself increases over time)

Based on the AIC values and through visual observation, we can see that both exponential and quadratic models provide better fits than the linear model.

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
```

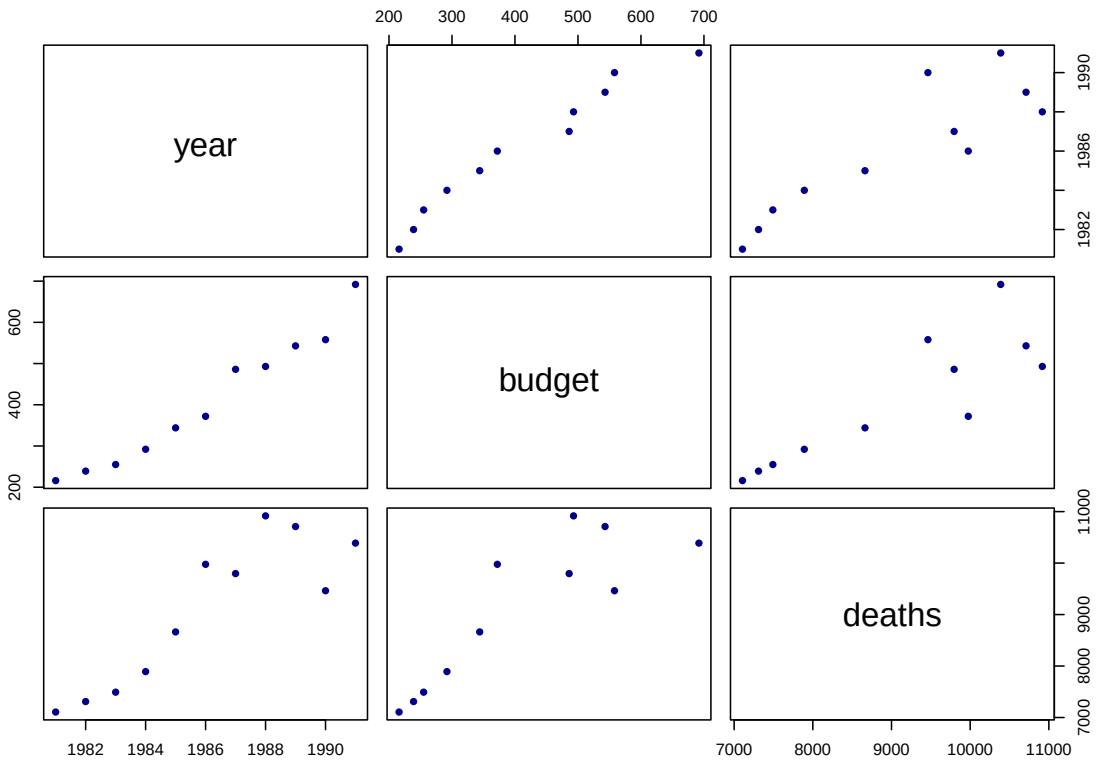
```r
# Compare models using AIC
AIC(linear_model_deaths, quadratic_model_deaths, exp_model_deaths)
```

```
                       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.

Computing Regression Lines

Now I'll compute the regression lines requested in the assignment. This will help us quantify the relationships we've been visualizing.

Regression Line: Deaths versus DEA Budgets

First, let's compute the regression line between drug deaths and DEA budgets:

```
# Compute regression for deaths vs budget
deaths_budget_reg <- lm(deaths ~ budget, data = dea_data)

# Display regression summary
summary(deaths_budget_reg)
```

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

Residuals:
```

```
   Min     1Q Median     3Q    Max
-918.5 -429.7 -255.2  348.4 1197.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5840.853    671.785   8.695 1.13e-05 ***
budget         7.898      1.547   5.107 0.000639 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 761.5 on 9 degrees of freedom
Multiple R-squared:  0.7434,    Adjusted R-squared:  0.7149
F-statistic: 26.08 on 1 and 9 DF,  p-value: 0.0006394
```

```r
# Extract coefficients for the equation
deaths_budget_coef <- coef(deaths_budget_reg)
deaths_budget_eq <- paste0("Deaths = ", round(deaths_budget_coef[1], 2),
                           " + ", round(deaths_budget_coef[2], 3), " × Budget")

# Print the equation
cat("Regression equation for Deaths vs Budget:\n", deaths_budget_eq)
```

```
Regression equation for Deaths vs Budget:
 Deaths = 5840.85 + 7.898 × Budget
```

We can interpret this regression as follows: for each additional million dollars in the DEA budget, there's an estimated increase of about 7.898 drug-induced deaths. The intercept suggests that with a theoretical budget of $0, we would expect 5841 deaths. However, this is clearly an extrapolation far beyond our data and lacks practical meaning.

Regression Line: Deaths versus Year

Next, let's compute the regression line between drug deaths and year:

```r
# We already computed this as linear_model_deaths, but let's display it more clearly
deaths_year_coef <- coef(linear_model_deaths)
deaths_year_eq <- paste0("Deaths = ", round(deaths_year_coef[1], 2),
                         " + ", round(deaths_year_coef[2], 2), " × Year")

# Print the equation and summary
cat("Regression equation for Deaths vs Year:\n", deaths_year_eq)
```

```
Regression equation for Deaths vs Year:
 Deaths = -746680.4 + 380.54 × Year
```

```r
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
```

This regression indicates that each year, drug-induced deaths increased by approximately 380.54 on average during the period 1981-1991.

Regression Line: DEA Budgets versus Year

Finally, let's compute the regression line between DEA budgets and year:

```r
# We already computed this as linear_model, but let's display it more clearly
budget_year_coef <- coef(linear_model)
budget_year_eq <- paste0("Budget = ", round(budget_year_coef[1], 2),
                         " + ", round(budget_year_coef[2], 2), " × Year")

# Print the equation and summary
cat("Regression equation for Budget vs Year:\n", budget_year_eq)
```

```
Regression equation for Budget vs Year:
 Budget = -91020.04 + 46.04 × Year
```

```r
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
```

The regression shows that the DEA budget increased by approximately $46.04 million per year during this period.

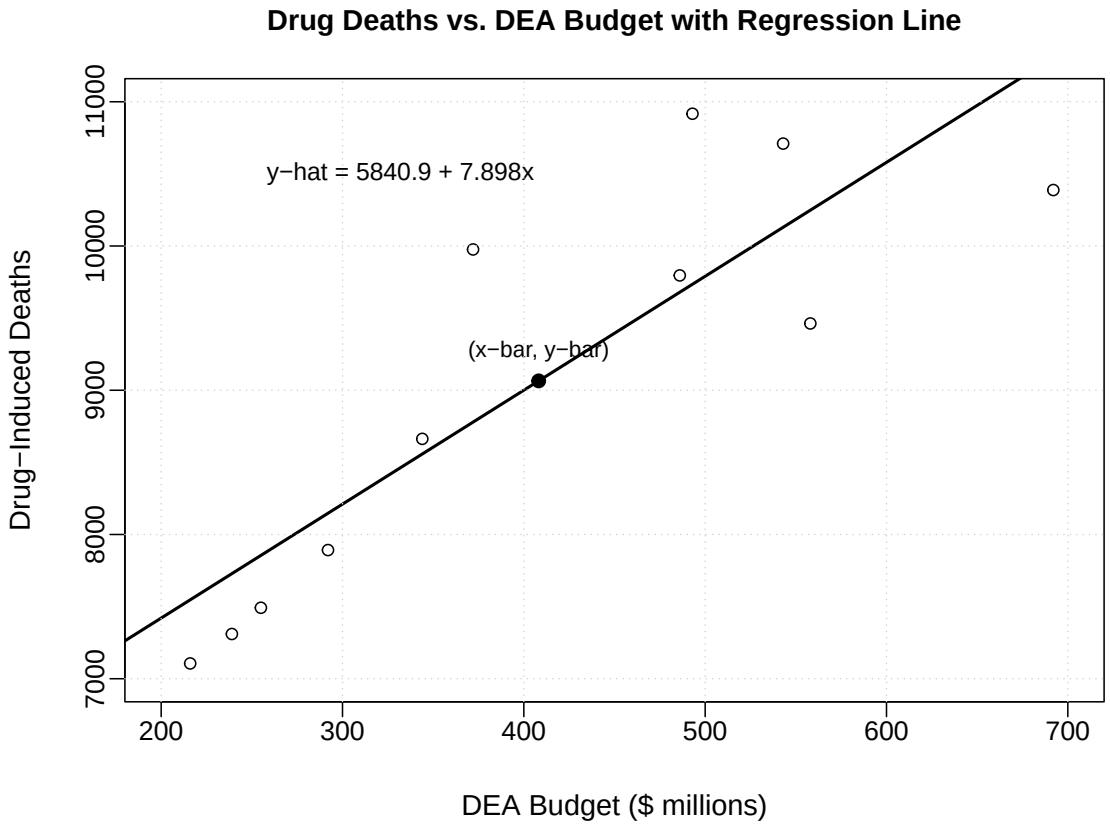
Scatter Plots with Regression Lines

Now let's create the three scatter plots with their respective regression lines, formatted similarly to Figure 12.3.3 in the textbook.

Scatter Plot 1: Deaths versus DEA Budget

```r
# Set plot parameters for a clean look
par(mar = c(5, 5, 4, 2), cex.lab = 1.2, cex.axis = 1.1, mgp = c(3, 0.5, 0))

# Create the scatter plot
plot(dea_data$budget, dea_data$deaths,
     main = "Drug Deaths vs. DEA Budget with Regression Line",
     xlab = "DEA Budget ($ millions)",
     ylab = "Drug-Induced Deaths",
```

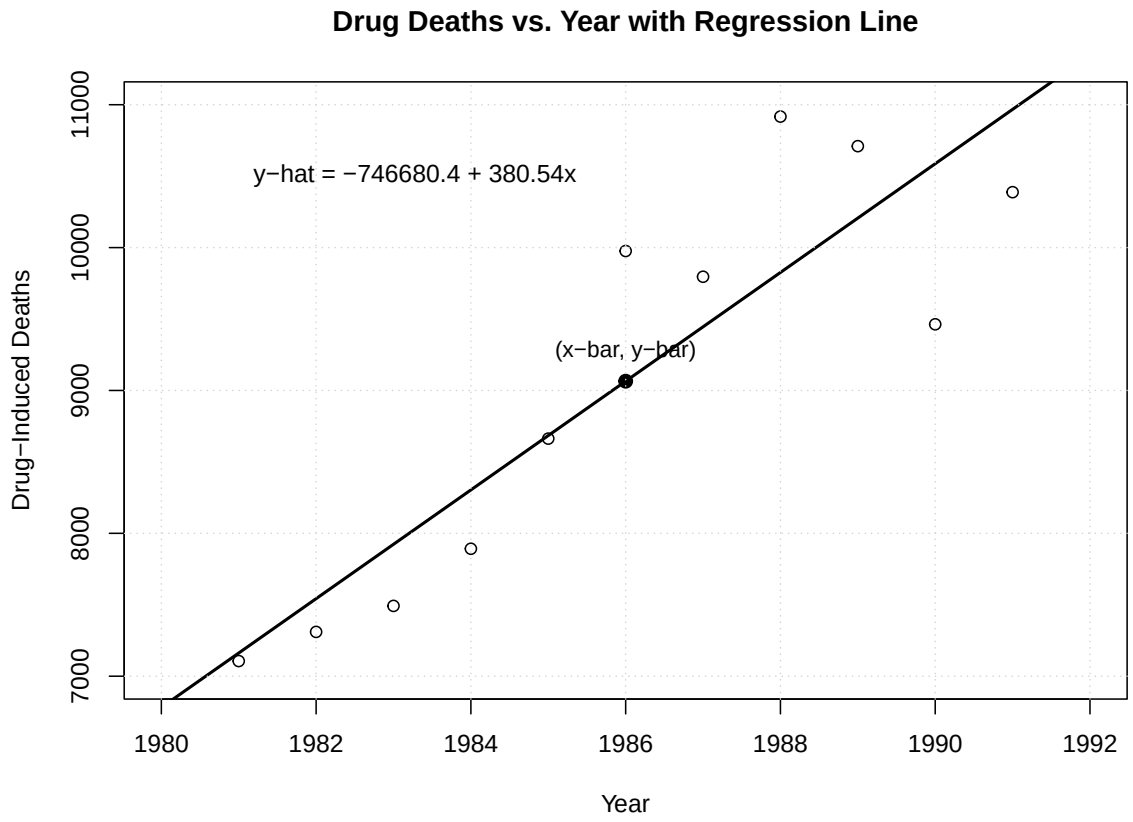
```
     pch = 1,  # Open circles like in Fig 12.3.3
     col = "black",
     xlim = c(200, 700),
     ylim = c(7000, 11000))

# Add regression line
abline(deaths_budget_reg, col = "black", lwd = 2)

# Add text showing the regression equation - using ASCII-safe characters
text(250, 10500,
     paste0("y-hat = ", round(deaths_budget_coef[1], 1), " + ",
            round(deaths_budget_coef[2], 3), "x"),
     pos = 4, cex = 1)

# Add text showing the mean point
mean_budget <- mean(dea_data$budget)
mean_deaths <- mean(dea_data$deaths)
points(mean_budget, mean_deaths, pch = 19, cex = 1.2)
text(mean_budget, mean_deaths, "(x-bar, y-bar)", pos = 3, cex = 0.9, offset = 0.8)

# Add intercept point
points(0, deaths_budget_coef[1], pch = 19, cex = 1.2)
text(0, deaths_budget_coef[1], "b0", pos = 4, cex = 0.9)

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



Scatter Plot 2: Deaths versus Year

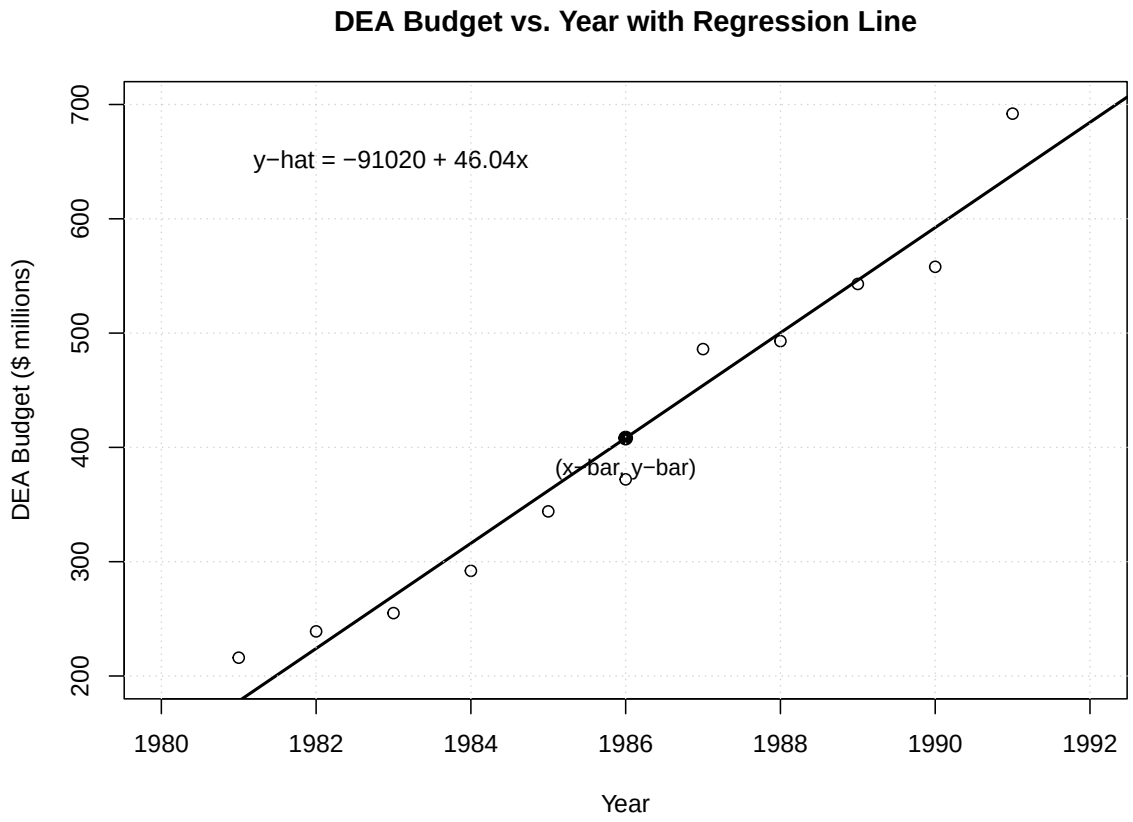
```
# Create the scatter plot for deaths vs year
plot(dea_data$year, dea_data$deaths,
     main = "Drug Deaths vs. Year with Regression Line",
     xlab = "Year",
     ylab = "Drug-Induced Deaths",
     pch = 1,
     col = "black",
     xlim = c(1980, 1992),
     ylim = c(7000, 11000))

# Add regression line
abline(linear_model_deaths, col = "black", lwd = 2)
```

```r
# Add text showing the regression equation - using ASCII-safe characters
text(1981, 10500,
     paste0("y-hat = ", round(deaths_year_coef[1], 1), " + ",
            round(deaths_year_coef[2], 2), "x"),
     pos = 4, cex = 1)

# Add text showing the mean point
mean_year <- mean(dea_data$year)
points(mean_year, mean_deaths, pch = 19, cex = 1.2)
text(mean_year, mean_deaths, "(x-bar, y-bar)", pos = 3, cex = 0.9, offset = 0.8)

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



Scatter Plot 3: DEA Budget versus Year

```
# Create the scatter plot for budget vs year
plot(dea_data$year, dea_data$budget,
     main = "DEA Budget vs. Year with Regression Line",
     xlab = "Year",
     ylab = "DEA Budget ($ millions)",
     pch = 1,
     col = "black",
     xlim = c(1980, 1992),
     ylim = c(200, 700))

# Add regression line
abline(linear_model, col = "black", lwd = 2)

# Add text showing the regression equation - using ASCII-safe characters
text(1981, 650,
     paste0("y-hat = ", round(budget_year_coef[1], 1), " + ",
            round(budget_year_coef[2], 2), "x"),
     pos = 4, cex = 1)

# Add text showing the mean point
mean_budget <- mean(dea_data$budget)
points(mean_year, mean_budget, pch = 19, cex = 1.2)
text(mean_year, mean_budget, "(x-bar, y-bar)", pos = 1, cex = 0.9, offset = 0.8)

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



Extended Statistical Analysis and Findings

Let's examine our findings more comprehensively to understand what our regression analysis tells us about the relationships between these variables.

Comparing R-squared Values

```
# Extract R-squared values
r2_deaths_budget <- summary(deaths_budget_reg)$r.squared
r2_deaths_year <- summary(linear_model_deaths)$r.squared
r2_budget_year <- summary(linear_model)$r.squared

# Display comparison
r2_df <- data.frame(
```

```
  Relationship = c("Deaths vs Budget", "Deaths vs Year", "Budget vs Year"),
  R_squared = c(r2_deaths_budget, r2_deaths_year, r2_budget_year)
)

knitr::kable(r2_df,
             col.names = c("Relationship", "R-squared"),
             caption = "Comparison of R-squared Values",
             digits = 3)
```

Table 3: Comparison of R-squared Values

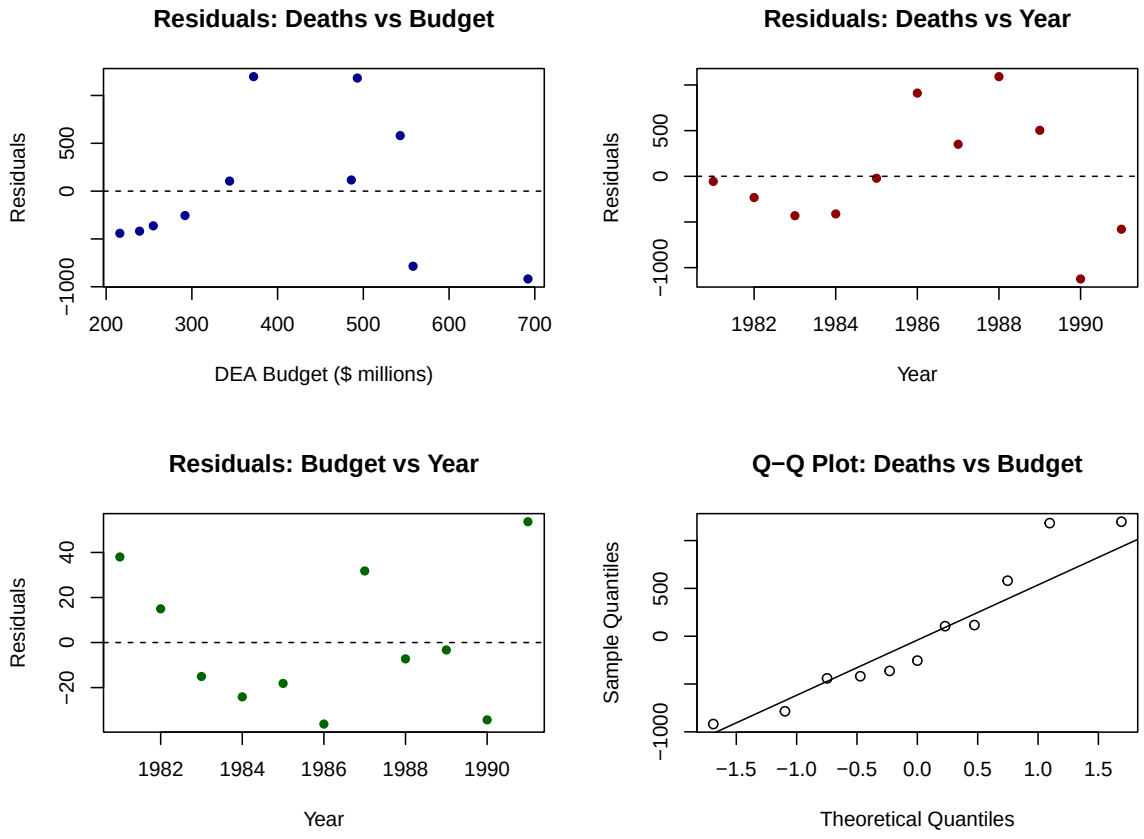
Relationship	R-squared
Deaths vs Budget	0.743
Deaths vs Year	0.783
Budget vs Year	0.962

Residual Analysis

Let's examine the residuals to check how well our linear models fit:

```
# Calculate residuals for all models
residuals_deaths_budget <- residuals(deaths_budget_reg)
residuals_deaths_year <- residuals(linear_model_deaths)
residuals_budget_year <- residuals(linear_model)

# Plot residuals for Deaths vs Budget
par(mfrow = c(2, 2))
plot(dea_data$budget, residuals_deaths_budget,
     main = "Residuals: Deaths vs Budget",
     xlab = "DEA Budget ($ millions)",
     ylab = "Residuals",
     pch = 16,
     col = "darkblue")
abline(h = 0, lty = 2)

# Plot residuals for Deaths vs Year
plot(dea_data$year, residuals_deaths_year,
     main = "Residuals: Deaths vs Year",
     xlab = "Year",
     ylab = "Residuals",
```

```r
     pch = 16,
     col = "darkred")
abline(h = 0, lty = 2)

# Plot residuals for Budget vs Year
plot(dea_data$year, residuals_budget_year,
     main = "Residuals: Budget vs Year",
     xlab = "Year",
     ylab = "Residuals",
     pch = 16,
     col = "darkgreen")
abline(h = 0, lty = 2)

# QQ plot for one of the models to check normality
qqnorm(residuals_deaths_budget, main = "Q-Q Plot: Deaths vs Budget")
qqline(residuals_deaths_budget)
```



```r
# Reset plot settings
par(mfrow = c(1, 1))
```

Final Interpretations and Conclusions

After computing the regression lines and performing a more detailed analysis, I can now provide more concrete findings about the relationships between drug deaths, DEA budget, and time:

1. **Deaths vs DEA Budget**:
    - The regression equation shows that deaths increase as budget increases
    - However, this doesn't mean budget causes deaths (correlation causation)
    - Both are influenced by time (the lurking variable)
    - This is evident from the residual plot which shows patterns rather than random scatter
2. **Deaths vs Year**:
    - Deaths generally increased over the period 1981-1991
    - The linear trend explains about 78.3% of the variation in deaths
    - There appears to be some non-linear pattern in the residuals, suggesting a more complex relationship
    - Some years (particularly 1989-1991) show notable deviations from the trend
3. **Budget vs Year**:
    - The DEA budget shows a strong linear relationship with time
    - The R-squared value of 0.962 indicates that a large portion of budget variation is explained by time
    - There's a clear upward trend in funding over this period
    - The residual pattern suggests a potential non-linear trend (as we saw earlier with quadratic and exponential models)
4. **Overall Conclusion**:
    - The strong relationship between Deaths and Budget appears to be primarily due to both increasing over time
    - There's no evidence from this analysis that increased DEA funding reduced drug deaths
    - In fact, despite budget increases, deaths continued to rise during this period
    - This suggests that the drug crisis during this period was not effectively addressed by simply increasing enforcement funding
    - The data shows a classic example of how lurking variables (time in this case) can create apparent correlations between unrelated variables

- Both policy makers and statisticians need to be careful about assuming causative relationships when multiple variables change over time

This analysis reinforces the importance of identifying lurking variables and avoiding spurious correlations in data analysis. When multiple variables increase over time, they will inevitably show correlation even without any causal relationship.