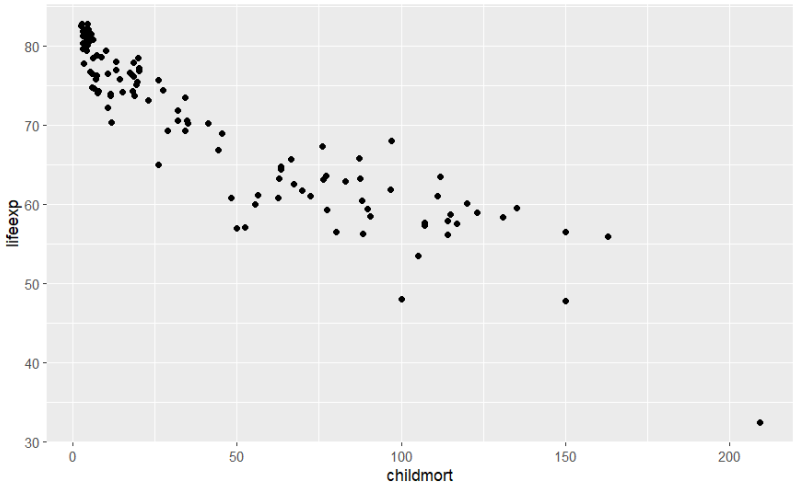
Mahek Patel, Alisa Prinyarux, Sharon Feinleib, Tania Ommer, Sneha Augustine
01:960:291 Statistical Inference for Data Science
Dr. Zhao
Final Project

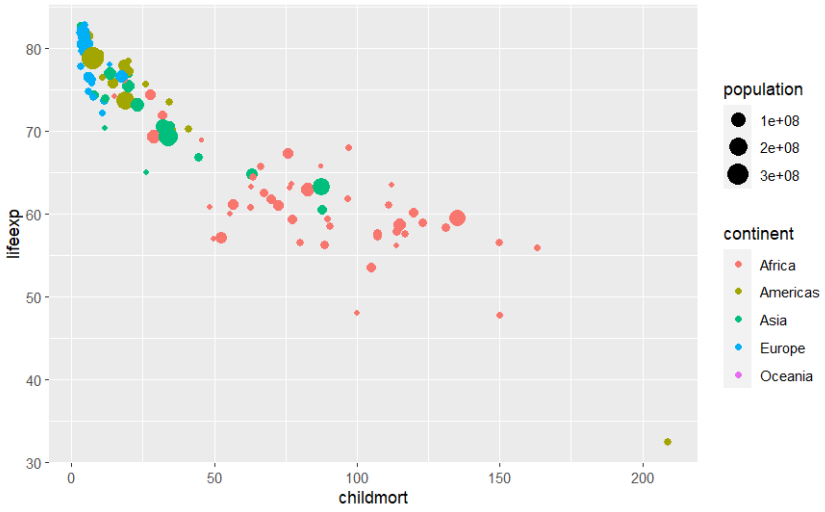
Exploratory Data Analysis

The all dataset provided to the class was used for this report.

The first step in this project was to explore associations between variables. We first created a scatterplot to see if there is a significant, general association between child mortality and life expectancy in 2009. There **is** a moderate negative linear relationship between the two. As child mortality decreases, life expectancy increases (see Graph 1). This is the general association with no other variables considered. We then investigated life expectancy rates among the different continents. When child mortality rates are plotted against life expectancies with population sizes and continents factored in, European and Oceanic countries have the highest life expectancies (see Graph 2). Data concerning Oceania have the least variability. As plots were created, we discovered that the variables murder, pop, childmort, and co2 are significant, i.e. there is a noticeable association and life expectancy. Plots are not the only part of the situation. The regression models themselves have to be analyzed and considered.



Graph 1



Graph 2

Multicollinearity Analysis

- Model 3:

In the first regression model (Model 3), we included population, co2, murder, and child mortality as predictor variables to explain life expectancy. The Variance Inflation Factor (VIF) values for these variables suggests less collinearity with the highest VIF observed for the childmort variable at 1.63. Despite this collinearity, the model demonstrates a high R-squared value of 0.849, suggesting that these variables collectively explain a substantial proportion of the variance in life expectancy.

```
> ##model 3
> model3 <- lm(lifeexp ~ population+ co2 + murder+ childmort, data = all)
> get_regression_summaries(model3)
# A tibble: 1 × 9
  r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
      <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
1     0.849         0.843  14.4  3.80  3.88      154.       0     4   115
> vif(model3)
population        co2     murder  childmort
  1.535023   1.655502   1.518227   1.626713
```

- Model 5:

The second model (Model 5) introduced interaction terms, specifically the interaction between co2 and population, along with additional variables such as water, murder, and popdensity. The model's R-squared value of 0.65 suggests a moderate level of explanatory power, explaining approximately 65% of the variance in life expectancy. Upon evaluating multicollinearity using Variance Inflation Factors (VIF), the results indicate acceptable levels of collinearity for the individual predictors. The highest VIF is observed for the interaction term co2:population, with a value of 2.40, which is within an acceptable range.

```
> model5<- lm(lifeexp ~ co2 * population + water + murder + popdensity , data = all)
> get_regression_summaries(model5)
# A tibble: 1 × 9
  r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
      <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
1      0.65          0.63  33.4  5.78  5.96      33.4       0     6   115
> vif(model5)
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
          co2     population          water         murder     popdensity co2:population
     1.989724       2.556352       1.530878       1.537338       1.038160       2.400935
```

- Model 6:

Model 6 extended the complexity by incorporating interaction terms (co2:population) along with health spend, water, murder, and continent. The VIF values for this model indicated moderate collinearity with the highest VIF observed for the interaction term co2:population at 3.09. While there is collinearity, it is not as severe as in Model 7. Despite this, the model shows an R-squared value of 0.713, suggesting an improvement in explanatory power.

```
> model6 <- lm(lifeexp ~ co2*population + healthspend + co2 + water+ murder + continent,
+              data = all)
> get_regression_summaries(model6)
# A tibble: 1 × 9
  r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
      <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
1     0.713         0.686  27.3  5.23  5.50      25.9       0    10   115
> vif(model6)
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
                   GVIF Df GVIF^(1/(2*Df))
co2            3.586558  1        1.893821
population     3.078830  1        1.754659
healthspend    3.064096  1        1.750456
water          2.307947  1        1.519193
murder         1.713844  1        1.309139
continent      5.201027  4        1.228885
co2:population 3.085684  1        1.756611
```

- Model 7:

Model 7 incorporates second-order terms and interactions, introducing a quadratic term for child mortality (I(murder^2)) along with various other variables, such as gdpcapita, population, co2, water, popdensity, murder, and continent. The R-squared value for this model is 0.72, indicating that approximately 72% of the variability in life expectancy is explained by the included variables. The VIF results reveal notable collinearity concerns for certain predictors. The variable I(murder^2), representing the square of murder rates, exhibits a VIF of 8.18, indicating a high level of collinearity. This suggests that the inclusion of the squared term for murder rates is leading to a substantial correlation with other predictors in the model. Similarly, the variable murder has a high VIF of 10.50, suggesting strong collinearity as well.

```
> model7 <- lm(lifeexp ~ I(murder^2)+ gdpcapita + population + co2 + water + popdensity +
+                murder + continent, data = all)
> get_regression_summaries(model7)
# A tibble: 1 × 9
  r_squared adj_r_squared   mse  rmse sigma statistic p_value    df  nobs
      <dbl>         <dbl> <dbl> <dbl> <dbl>     <dbl>   <dbl> <dbl> <dbl>
1      0.72         0.691  26.7  5.16  5.46      24.1       0    11   115
> vif(model7)
                 GVIF Df GVIF^(1/(2*Df))
I(murder^2)  8.180189  1        2.860103
gdpcapita    3.264023  1        1.806661
population   1.884795  1        1.372878
co2          3.190797  1        1.786280
water        2.310124  1        1.519909
popdensity   1.186784  1        1.089396
murder      10.497305  1        3.239954
continent    4.617786  4        1.210749
```

Collinearity plays a crucial role in the interpretability of regression models. The VIF values provide insights into the potential impact of collinearity on the precision of coefficient estimates. Generally, VIF values below 5 (the lower the better) are considered acceptable, and in models 3 and 5, all predictors fall within this range. The interpretation of the model's coefficients is thus less likely to be compromised due to multicollinearity. While moderate collinearity may not severely impact the models' performances, high collinearity, as observed in Model 6 and Model 7, requires careful consideration. In the presence of severe collinearity, we attempted model simplification and tested multiple predictors to generate the four models above since these yielded high R-squared values.

Regression tables

Model 3

term <chr>	estimate <dbl>	std_error <dbl>	statistic <dbl>	p_value <dbl>	lower_ci <dbl>	upper_ci <dbl>
intercept	77.140	0.879	87.787	0.000	75.399	78.882
population	0.000	0.000	1.001	0.319	0.000	0.000
co2	0.256	0.109	2.342	0.021	0.039	0.472
murder	0.000	0.000	-1.489	0.139	0.000	0.000
childmort	-0.182	0.010	-17.932	0.000	-0.202	-0.162

Model 5

term <chr>	estimate <dbl>	std_error <dbl>	statistic <dbl>	p_value <dbl>	lower_ci <dbl>	upper_ci <dbl>
intercept	11.099	6.351	1.748	0.083	-1.489	23.687
co2	0.773	0.184	4.202	0.000	0.408	1.138
population	0.000	0.000	0.806	0.422	0.000	0.000
water	0.609	0.073	8.336	0.000	0.464	0.754
murder	0.000	0.000	-0.879	0.381	0.000	0.000
popdensity	0.000	0.001	0.320	0.749	-0.001	0.002
co2:population	0.000	0.000	-1.194	0.235	0.000	0.000

Model 6

term <chr>	estimate <dbl>	std_error <dbl>	statistic <dbl>	p_value <dbl>	lower_ci <dbl>	upper_ci <dbl>
intercept	27.076	6.916	3.915	0.000	13.360	40.791
co2	0.321	0.228	1.408	0.162	-0.131	0.772
population	0.000	0.000	0.603	0.548	0.000	0.000
healthspend	0.001	0.000	1.787	0.077	0.000	0.002
water	0.404	0.083	4.884	0.000	0.240	0.568
murder	0.000	0.000	-0.877	0.382	0.000	0.000
continent: Americas	6.259	1.891	3.310	0.001	2.509	10.009
continent: Asia	4.756	1.922	2.475	0.015	0.945	8.566
continent: Europe	7.051	2.080	3.390	0.001	2.926	11.176
continent: Oceania	6.665	4.643	1.435	0.154	-2.543	15.873
co2:population	0.000	0.000	-1.037	0.302	0.000	0.000

Model 7

term <chr>	estimate <dbl>	std_error <dbl>	statistic <dbl>	p_value <dbl>	lower_ci <dbl>	upper_ci <dbl>
intercept	66.062	6.019	10.976	0.000	54.126	77.999
I(childmort^2)	-0.001	0.000	-10.559	0.000	-0.001	-0.001
gdpcapita	0.000	0.000	3.312	0.001	0.000	0.000
population	0.000	0.000	1.148	0.254	0.000	0.000
co2	0.005	0.148	0.036	0.972	-0.288	0.299
water	0.020	0.068	0.294	0.770	-0.114	0.154
popdensity	0.001	0.001	1.535	0.128	0.000	0.002
murder	0.000	0.000	-1.958	0.053	0.000	0.000
continent: Americas	7.227	1.306	5.535	0.000	4.637	9.816
continent: Asia	3.270	1.286	2.543	0.012	0.719	5.821
continent: Europe	6.924	1.410	4.911	0.000	4.128	9.720
continent: Oceania	8.313	3.167	2.625	0.010	2.033	14.594

Runner-ups vs Final Model

Model 3:

- R-squared (rsq): 0.849, Adjusted R-squared (adj rsq): 0.843
- Model 3, which features predictors such as population, carbon dioxide emissions (co2), murder rates, and child mortality, demonstrates a high overall fit with the data. The R-squared and adjusted R-squared values indicate that approximately 84.3% of the variability in life expectancy is explained by the model, considering the number of predictors.

Model 5 (Interaction Model):

- R-squared (rsq): 0.65, Adjusted R-squared (adj rsq): 0.63
- Model 5, which incorporates interaction terms involving co2 and population along with additional predictors like water, popdensity and murder, exhibits a lower overall fit compared to Model 3. The R-squared and adjusted R-squared values suggest that around 69% of the variability in life expectancy is explained by this model.

Model 6:

- R-squared (rsq): 0.713, Adjusted R-squared (adj rsq): 0.686

- Model 6, comprising co2, population, health spending, water, murder rates, and continent, shows a moderate overall fit. The R-squared and adjusted R-squared values indicate that approximately 71.3% of the variability in life expectancy is captured by this model.

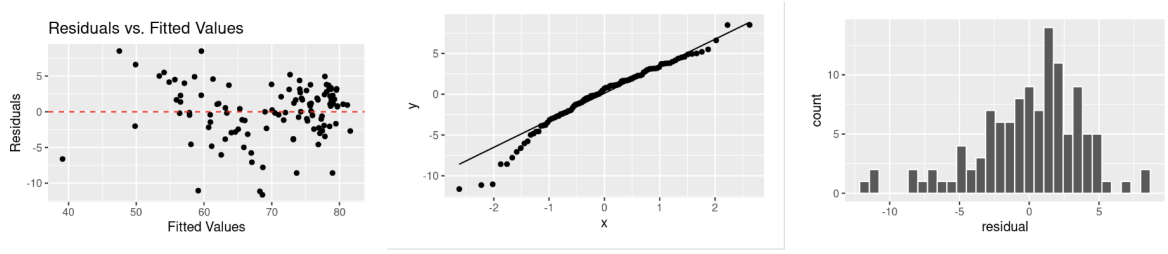
Model 7:

- R-squared (rsq): 0.72, Adjusted R-squared (adj rsq): 0.691
- Model 7, featuring squared murder rates, gdpcapita, popdensity, and other predictors, demonstrates a relatively strong overall fit. The R-squared and adjusted R-squared values suggest that approximately 72% of the variability in life expectancy is explained by this model.

Considering the goodness of fit and having been adjusted for model complexity, Model 3 stands out with the highest R-squared and adjusted R-squared values. This indicates that it provides the best balance between explaining variability and avoiding overfitting. However, it's essential to consider other factors, such as the assumptions of linear regression and the interpretability of the model, before making a final decision.

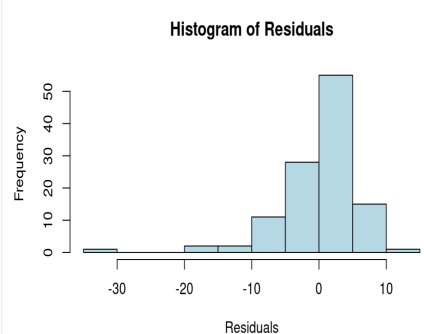
Residual Analysis

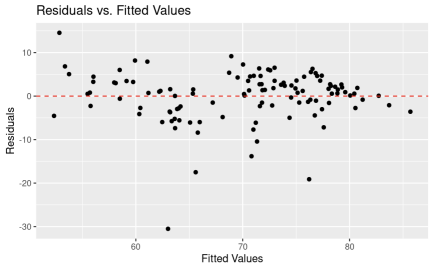
Model 3

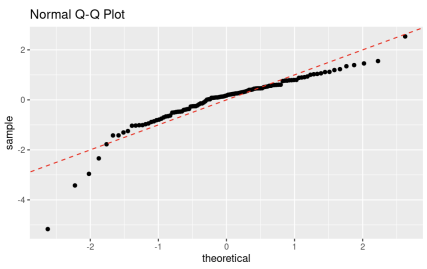


The constant variance and linearity is basically satisfied; the qq plot displays the residuals remaining near the regression line, but there are a few outliers and seems to be a curve that opens down in the residual plot. The r^2 is 0.849 meaning it has a relatively strong relationship since it is pretty close to 1.The histogram shows a semi-normal distribution, skewed left, with visible outliers. Independence is basically satisfied because the residuals are not dependent on one another.

Model 5







Looking at the plots, we can see that the constant variance and linearity is satisfied. The qq plot displays the residuals remaining near the regression line, but there are a few outliers. The r^2 is 0.6496416 meaning it has a relative relationship since it is not quite close to 1.The histogram shows a semi-normal distribution. However, it is very skewed to the left with visible outliers. By looking at the point, we can conclude that independence is satisfied because the residuals are not dependent on one another.

AIC and BIC in Model Selection

AIC(model3)
AIC(model5)
AIC(model6)
AIC(model7)

BIC(model3)
BIC(model5)
BIC(model6)
BIC(model7)

Model 3:

- Predictors: population, co2, murder, childmort
- AIC: 624.2791, BIC: 646.2386

Model 5 (Interaction Model):

- Predictors: co2, population, water, density, murder, popdensity
- AIC: 745.8474, BIC: 767.8069

Model 6:

- Predictors: co2, population, healthspend, water, murder, continent
- AIC: 730.8233, BIC: 763.7625

Model 7:

- Predictors: murder^2, gdpcapita, population, co2, water, popdensity, murder, continent
- AIC: 729.8945, BIC: 765.5786

Comparing models:

- For AIC, lower values are better. In this case, Model 3 has the lowest AIC, suggesting it might be the best-fitting model among those considered.
- For BIC, similarly, lower values are better. Again, Model 3 has the lowest BIC.

The models under consideration present a diverse array of predictor variables and interactions. Each model attempts to capture the complexity of the relationship between life expectancy and various factors. Model 3 includes essential variables such as population, carbon dioxide

emissions (co2), murder rates, and child mortality. Despite its simplicity, this model boasts the lowest AIC and BIC values, suggesting superior fit compared to the alternatives. In contrast, Model 5 introduces interaction terms involving co2 and population, murder, popodensity and water. Although the R-squared value is commendable at 0.65, the elevated AIC and BIC values indicate potential overfitting likely due to the intricate interactions introduced.

Model 6, featuring interactions between co2 and population and variables like health spending, water, murder rates, and continent, yields a competitive R-squared of 0.713. However, its AIC and BIC values fall between Model 3 and Model 5, indicating a trade-off between fit and model complexity. Model 7 introduces a polynomial term (murder^2) and additional variables such as gdpcapita and popdensity. Despite the complexity, the AIC and BIC values for Model 7 are marginally higher than those of Model 3, making it a reasonable alternative.

In conclusion, while Model 3 exhibits the best trade-off between goodness of fit and model simplicity, Models 5, 6, and 7 introduce complexities that do not significantly improve explanatory power. The introduction of interaction terms and higher-order polynomials seems to contribute to overfitting, potentially limiting the models' generalizability to new data. The decision on the "best" model should consider not only statistical measures but also theoretical soundness, interpretability, and the ability to generalize to new observations. Therefore, Model 3 emerges as the preferred model in this context, offering a parsimonious representation of life expectancy determinants without sacrificing predictive accuracy.

The final model is Model 3.