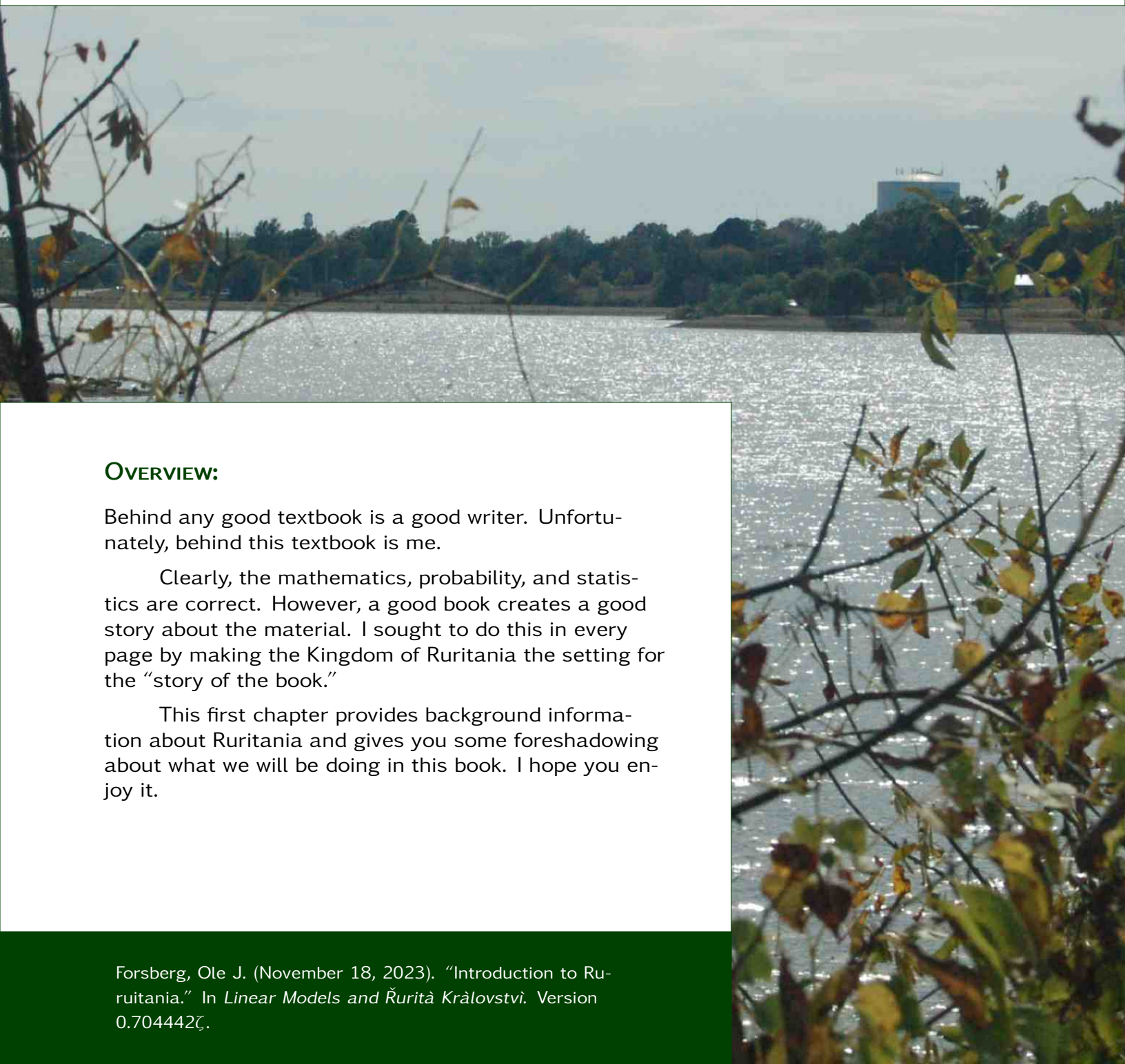
Chapter 1:
Introduction to Ruruitania



Overview:

Behind any good textbook is a good writer. Unfortunately, behind this textbook is me.

Clearly, the mathematics, probability, and statistics are correct. However, a good book creates a good story about the material. I sought to do this in every page by making the Kingdom of Ruritania the setting for the "story of the book."

This first chapter provides background information about Ruritania and gives you some foreshadowing about what we will be doing in this book. I hope you enjoy it.

Forsberg, Ole J. (November 18, 2023). "Introduction to Ruruitania." In *Linear Models and Řurità Kràlovstvì*. Version 0.704442ζ.

Chapter Contents

❧ ❧ ❧

You are about to undertake an educational journey. This journey will take you to new and exciting places. Here, you will turn your mathematical knowledge into statistical knowledge, giving you skills in detecting relationships between variables in life.

Figure 1.1: *His Majesty Rudolph II, King of Ruritania.*

This introduction has two primary purposes. The first is to introduce you to the Kingdom of Ruritania (Řurità Kràlovstvì). This fictional country is used as the backdrop for many of the examples. Why use Ruritania? It offers no conflicting information, thus allowing us to focus on the research questions at hand.

Second, this chapter provides many instances of analyses discussed in this text. Think of this chapter as a peek into the future, as a foreshadowing of the great things to come!

1.1: Background of Ruritania

The Kingdom of Ruritania (officially: Řurità Kràlovstvì) is a small kingdom (62 mi^2; 161 km^2) surrounded by Germany and the Czech Republic. Its inhabitants are Slavic-speaking Roman Catholics currently under the absolute monarchy of Rudolph II, pictured at right.

Under Rudolph II, ascension talks between Ruritania and the European Union have stalled. The two issues are the deficiency of democratic structures and an excess of controls on the press. Regardless of not being a member of the European Union, or of its monetary union (euro area), Rudolf pegged the Ruritanian Crown (*Koruna Řuritè*) to the Euro at 2.00Kř to €1.00. This offers greater stability for the koruna in the world currency market. Three decades ago, economic reform allowed the koruna to be entirely convertible on the world market. It also completely eliminated the black market in Ruritania.

1.1.1 The Geography

Ruritania is land-locked. It is located between southern Germany (Saxony) and western Czech Republic (Bohemia). The land is a beautiful combination of high mountains in the west (the Ruritanian Alps) and rolling farmland in the east (the Ruritanian Veldt).

1.1.2 The Government

Note that Ruritania is an autocratic kingdom, not a constitutional monarchy; the king rules as he sees fit. The king is advised by a council of ministers that he selects. These ministers need not be citizens of Ruritania. The current council is composed of five Ruritanians and one Oregonian: the current Minister of Economics, who is Knox-educated and from Portland, Oregon. These councilors also perform the tasks of a 'Committee King' when the king is unable to perform his duties as Chief of State and Head of Government.

Strešlau, which serves as the official capital city, lies on the rail line between Dresden and Prague and has a population of 9,313. The royal capital, Sčwånstein, has a population of 2,685 and lies on a spur (branch) line. According to the 2010 census, the total population of Ruritania is 15,295, with about 90% of the population living in its three main cities: Strešlau, Sčwånstein, and Děčín.

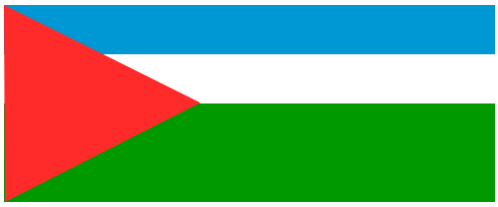


Figure 1.2: *The Vlajka, the current flag of the Kingdom of Ruritania adopted in 1954. Blue represents the sky; white, the snow-capped peaks; green, the lush farmland; and red, the passion of the people.*

Ruritania is divided into seven states (*stâty*). Each *stât* is named after its largest city. Thus, the seven *státy* are Děčín, Hora, Reka, Sčwånstein,

Strešlau, Venkovský, and Zámek. Each *státy* is further divided into five counties (*kraj*). Officially, the name of each *kraj* is the *stát* name followed by a letter between A and E.

1.2: Economics

While Ruritania's size limits its ability to diversify its economy, Rudolf has been an able administrator and businessman. As such, he was able to lift many Ruritanians out of poverty. When he took power in 1940, the GDP per capita (PPP) was approximately USD50, with a poverty rate of 99%. Today, it is approximately USD55,000 (the eighth highest in the world), with a poverty rate of just 50%.

The primary source of revenue for Ruritania is its banking industry — the source of 75% of its GDP (see Figure 1.3). The remainder comes from tourism (15%) and agriculture (10%). Ruritania's tourism industry primarily bases itself on winter recreation in the west. Because of favorable visa requirements, low costs for hotels, and fantastic skiing in its alps, Ruritania is the destination of choice for vacationers from (in descending order of visitors) the Czech Republic, Germany, Turkey, Oman, Morocco, United States, and Palau.

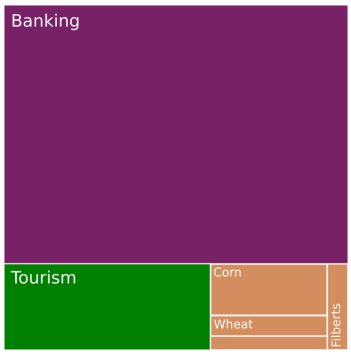


Figure 1.3: *A tree diagram of the economic output of Ruritania.*

The primary crops are corn (55%), summer wheat (25%), and filbert nuts (15%), with soybeans and hops being secondary. What corn is not eaten is exported to Germany (75%) and the Czech Republic (25%). Similar export patterns hold for the excess wheat. Filbert exports go to the Czech Republic (40%), Germany (35%), and Switzerland (25%), where they are turned into delicious confections. Imports to Ruritania come from Russia, Turkey, and Oman (petroleum), and the Czech Republic and Germany (manufactured goods and foodstuffs).

Because of the strength of the monarchy, Ruritania is neither a production point nor a transshipping point for drugs. Illicit drug use is the lowest in Europe, with approximately 2% of the population using marijuana, and none using harder drugs.

1.3: US–Ruritanian Relations

The United States and Ruritania share full diplomatic recognition. However, the United States does not have an ambassador to Ruritania. The US interests in Ruritania are handled by the US Ambassador to Poland, Mark Brzezinski (since February 22, 2022). This is not an unusual circumstance, as embassies are quite expensive. Unfortunately, this reduces the amount of reliable information coming out of Ruritania.

Rudolf was a staunch ally of the United States during the Cold War. However, with the geographic position of his country in the world (entirely surrounded by Soviet satellite states East Germany and Czechoslovakia), he was unable to offer the United States anything other than intermittently vocal, moral support in the United Nations. For fear of losing sovereignty, Rudolf often kept quiet and followed the lead of the Soviet Union in all but domestic economic matters.

Figure 1.4: *US Ambassador to Poland, Mark Brzezinski.*

When the Soviet Union fell, Rudolf increased his vocal support of the United States and its efforts to bring peace and prosperity to the world. As Rudolf often pointed out to various US presidents, Ruritania has never seen one of its sons die in battle. With his declining health and advancing age, Rudolf became much vehement in his support of the United States, especially with respect to the Global War on Terror.

1.4: Illustrations of Analyses

That concludes the background to Ruritania. The following examples show you what you will be able to do by the time you finish this book. The illustrations provide neither the data nor the code. All they provide are examples of how linear models are helpful to Ruritania... and to the rest of the world. Read through them and become excited about what this term will bring to you!

1.4.1 Illustration: The Missing Kraj

Every 10 years of a king's reign, Ruritania holds a census. Their last was in 2010, marking King Rudolph's 70th year as monarch. After compiling all information, they discovered that information was missing for one *kraj* (county). This is unfortunate, because King Rudolph needs the information to evaluate his latest five-year plan and determine what he should do to make it work better.

To fill in the missing information, we can regress all other variables on the GKP per capita, then predict the GKP per capita based on the known values for the missing kraj.

With this data and model, we just predict the GKP per capita in the kraj. The most-likely value is $2400, with a 95% prediction interval from $2100 to $2650.

From this information, His Majesty concludes that the plan helped the entire country, but did a better job with the rural areas than the urban. As a result of this analysis, he asks his ministers to generate a plan that does a better job of spreading the prosperity to more of the Kingdom.

impute

This use of regression has a very important use: estimating values for missing data in a data set (single imputation). Frequently, the amount of missing data will be significant with respect to the amount of complete data. In such a case, the researcher may use multiple imputation to create multiple data sets, estimate the parameters of interest on each, and report them and their standard errors.

1.4.2 Illustration: Ruritanian Crops

As usual, His Majesty Rudolph II would like some input from his Council of Ministers on his next five-year plan. Currently, the primary crop in Ruritania is corn. To help optimize the profits made by farmers, Rudolph wants to know if that crop should be changed to summer wheat or to soybeans.

To help him, let us model the relationship between farmer profit and crop in Ruritania. The dependent variable is the profit per acre, and the independent variable is the crop. Using linear models, we see that it is more profitable at this point to grow wheat. The average profit per acre is $845.75. This is about $150 greater than that of corn and about $300 greater than soybeans.

Linear models can also be applied to cases where the independent variable is categorical, as here. This method is actually termed "analysis of variance" (ANOVA). The difference between ANOVA and regression is only conceptual, in that each level of the categorical variable is treated as a separate variable.

Using statistics to inform policy decisions is an important use. However, we professionals need to be aware of the limitations of our research — and let our clients know them as well. Here, we only looked at the current average profit per acre for each crop. Future prices may fluctuate enough to make soybeans more valuable. Furthermore, shifting all Ruritanian crops to wheat puts the entire economy at risk of a drop in wheat prices. Diversification may be the better strategy, with some of the profit shared among all farmers.

1.4.3 Illustration: Cows in Děčín

The voters of Děčín are being sent to the polls to vote on a city referendum that proposes to limit the number of cows that can be kept inside the city. The wording on this referendum is quite similar to ones proposed over the past several years.

Given the information from previous votes, the Director of the Independent Electoral Commission (NVK) estimates that the probability the referendum has of passing is 40%, with an estimated 48% of the voters supporting it.

❧ ❧ ❧

Note that this research question deals with the *probability* of winning, not just the best estimate of the vote in favor. This requires estimating the entire probability distribution of the dependent variable.

While there are some rather sophisticated methods, we will be able to answer a similar question using Monte Carlo simulation. Such simulation consists of drawing large samples from each of the parameter-estimate distributions, calculating a predicted outcome for each of those sets of estimates, and examining the distribution of these predictions.

Monte Carlo estimation is a powerful technique that allows you to estimate results when the assumptions of the mathematical model are not fully met by the data.

1.4.4 Illustration: Wealth in Ruritania

The gross domestic product (GDP) per capita is one of many measures of average wealth in countries. If

extant theory is correct, then the wealth in the country is directly affected by the level of corruption in the government — countries with higher levels of corruption should be poorer (on average) than those with low levels of corruption. Furthermore, if theory is correct, the level of democracy in a country should *also* influence the country's level of wealth — countries with higher levels of democracy should be wealthier than countries with lower levels of democracy.

His Majesty is curious to see how Ruritania fits in this model. If the actual GDP per capita is greater than what is expected from modeling the rest of the world, then Rudolph is doing a great job as king. Otherwise, he needs to improve the lot of his people.

And so, to help Rudolph, we predict that the GDP per capita for Ruritania, according to the model, is $26,795.64, with a 95% prediction interval from $5232 to $48,360. Since the actual GDP per capita is $55,000, King Rudolph is happy that he is better than average at guiding Ruritania forward towards prosperity.

❧ ❧ ❧

Frequently, we can use our models in novel ways. Usually, we would model the data and calculate predictions and confidence intervals.

However, if we have confidence in our model, we can use it to determine which units are under- or over-performing expectations (the line of best fit). In this case, Ruritania's GDP per capita is significantly higher than what the model predicts.

This means either the model needs to take more covariates into consideration or that Ruritania is much more prosperous than one would expect... or both.

1.4.5 Illustration: Elections in Ruritania

Even though it is an absolute monarchy, national elections are held in Ruritania to elect members of the Ruritanian parliament, the *Národní Shromáždění* (National Assembly).

After the most recent election, the exiles in Denmark claimed that the ballot boxes were stuffed. That is, the ballot boxes had votes for the government party in them even before voting began. Because guarantees of the "secret ballot" are built into the Ruritanian Constitution, the ballot boxes are opaque. As such, there is no direct evidence of stuffing.

Ordinary least squares regression is not well-suited for this type of data. There is inherent heteroskedasticity in the proposed model. However, we can use Binomial regression to test the election for evidence of ballot box stuffing.

As a result of our analysis, we were able to detect some evidence for stuffing (p-value = 0.0441). However, because the p-value is so high, we should not claim to have found a smoking gun. To claim something so heinous, we should really contemplate the real meaning of the p-value.

❧ ❧ ❧

Here, we had a good understanding of the data-generating process. This allowed us to use that understanding to create a stronger model. Heteroskedasticity can be "adjusted for" in ordinary least squares. It should be a full part of the model, however.

1.4.6 Illustration: Insurance in Ruritania

The decision to buy life insurance is related to several variables, including age and income. We would like to explore this relationship in Ruritania.

Since the dependent variable is dichotomous (life insurance purchased *or* life insurance not purchased), we need a new type of regression to ensure that our predictions make sense. One option is called logistic regression.

Using this regression, we find significant positive relationships between the person's age and the likelihood to buy life insurance, as well as between the person's income and the likelihood to buy life insurance.

Additionally, we predict that the Knox graduate on the Council of Ministers, a 65-year-old making $125,000 annually, has a 74% chance of having life insurance.

❧ ❧ ❧

In this example, we had to use a different type of regression because the dependent variable was dichotomous, could only take on two values. This type of regression is known collectively as logistic regression, even though the link function can be almost any that map the real line to the interval $(0,1)$. Such functions include the venerable logit function (inverse of the logistic function). It also includes the probit (used in a lot of medical studies) and the cauchit (used in some financial studies to allow for highly variable events).

It was this class of problems that forced Nelder and associates to formulate an over-arching framework for regression. He called it the "generalized linear model" (GLM). While he rues the name to this day, he created it to signify that this class of regression problems is actually just a generalization of the class that can be solved using ordinary least squares regression. In other words, OLS regression is a special case of GLM regression.

1.4.7 Illustration: Warmth for the King

Finally, let us help the King be more beloved of his people, if that is possible. We took a poll and asked Ruritanians their 'warmth of feeling' for Rudolph and his political agenda. In addition to this one variable, we also asked several demographic questions, allowing us to provide suggestions to the King. The demographic information includes the gender, the race, the age, and the number of years of education. The response variable has four ordered levels: Strongly Disagree, Disagree, Agree, and Strongly Agree.

With this information, we are able to let the King know that he is widely loved, but that women tend to agree with his policies more than do men. Furthermore, the better-educated also tend to support him more. The younger members of Ruritania also feel warmer towards him and his agenda. Finally, there was no relationship between race and support; all races seem to support him equally.

With this information from the poll, what can Rudolph do to help his people? Such is the question royals have asked for generations.

❧ ❧ ❧

Noting that the dependent variable is an ordinal variable, we could not use ordinary least squares regression. We had to use something called "ordinal regression." The concepts behind ordinal regression are quite similar to those behind other types of regression covered in this book. The mathematics are a bit more difficult, however.

Thankfully, statistical programs make using ordinal regression almost as easy as using other types. We just have to know how to get the data in the right form for the program *and* how to test the assumptions made by the technique.

In fact, this seems to be the lesson we need to learn throughout this course. The concepts are quite similar. The mathematics are different. And, to make usage easier, those who write statistical environments try to make the functions as similar as they can.

1.5: Conclusion

And that brings us to the end of this introductory chapter. In this chapter, you were introduced to the incredible Kingdom of Ruritania. Ruritania offers us a rich source of examples, which will be exploited throughout the text.

This chapter also offered several examples that foreshadow what you will learn in this course. While you may not be able to do — or understand the underlying theory for — any of them at this point, you will by the end of the course.

Figure 1.5: *His Majesty Rudolph II, King of Ruritania, in 1952.*

So, take a deep breath and turn the page to the book's first part: Ordinary Least Squares (OLS). In this part of the book, we start by asking what we mean by "summarizing the relationship with a line of best fit." From that point, we leverage that definition to inform our mathematics, thus allowing us to create formulas for estimating the population parameters.

After that chapter, we use the mathematics and elementary probability theory to create test statistics and confidence intervals for testing hypotheses of interest.

And after that... the sky is the limit! It is a great journey, and King Rudolph II thanks you for taking it.