JOURNAL OF STATISTICS AND DATA SCIENCE EDUCATION
2025, VOL. 00, NO. 0, 1–16
https://doi.org/10.1080/26939169.2025.2490027

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

Multiple Regression with Transformations and Variable Selection at the Industrial Scale

Gus Greivel[a,b], Soutir Bandyopadhyay[b], Alexandra M. Newman[a], and Brian G. Thomas[c]

[a]Operations Research with Engineering Graduate Program, Colorado School of Mines, Golden, CO; [b]Department of Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO; [c]Department of Mechanical Engineering, Colorado School of Mines, Golden, CO

ABSTRACT
This article explores the development of a statistical model to predict roll forces during hot rolling in a commercial steel mill based on a dataset from an industrial-scale operation; this dataset consists of 2255 individual coils, processed through five roll stands in the mill, for a total of 11,275 distinct data points. This work is informed by a physics-based equation and applies variable transformations, variable selection and model comparisons; it illustrates several topics introduced throughout a regression modeling course, including data visualization, multiple linear regression, residual diagnostics and heteroscedasticity, variable transformations, collinearity and confounding due to strong correlations between predictor variables, the false correlation paradox, variable selection, and measures of model quality for predictive use. Extensions, including autocorrelation, mixed models, and resampling methods, are also suitable as more advanced topics and for open-ended courses. Finally, we add an interesting real dataset and associated analysis to the publicly available resources for statistics and data science education.

ARTICLE HISTORY
Received March 2023
Accepted March 2025

KEYWORDS
Continuous steel casting; Model interpretation; Prediction; Transformations; Variable selection

1. Introduction

The Colorado School of Mines is a mid-size, public university specializing in engineering and applied sciences with an emphasis on balancing applied research and undergraduate instruction. Within the context of a larger optimization project for an industry sponsor, we have identified a statistical modeling question and associated dataset that we incorporate in our MATH324: Statistical Modeling course. The sponsor has agreed to allow this dataset to be used and shared with the broader statistics education community (with proprietary information removed from the raw data). The research project in question required that a model be constructed to better predict the force necessary in a steel rolling mill operation based on several available predictor variables. While working with a group of engineers (our clients) with a solid sense of the physical properties involved, we negotiated our way to a model with sound predictive properties and reasonable explainability from the perspective of the engineers in the room. We describe the process through the lens of a consultation process with the clients. As the process evolved, we noted that our consultations addressed multiple topics from our MATH324 course, and we are now incorporating this dataset into the corresponding lectures and activities.

For perspective, MATH324 is a sophomore and/or junior-level course at the Colorado School of Mines that has our Introduction to Statistics course as a prerequisite; we therefore assume an exposure to basic inferential statistics and some experience in **R**. Given the composition of the degree programs offered at Mines, we can also assume that students have taken three semesters of Calculus and one semester of Differential Equations, with an emphasis on modeling physical applications throughout. Students have taken at least one semester of Physics for Engineers, as well as an interdisciplinary Engineering Design course. Throughout our first- and second-year curriculum, all of our students have a broad preparation in STEM that emphasizes application and real-world contextualization. As a consequence, our students are enthusiastic about working with real-world data that intersects with interdisciplinary applications relevant to the broader campus.

MATH324 acts as an introduction to linear regression, with variable transformations, variable selection, inference, and validation included in the curriculum; the course is intended as an early exploration into statistical modeling, emphasizing hands-on exploration of real world datasets. It serves as a prerequisite to junior and/or senior-level courses on Time Series, Spatial Statistics, General Linear Models, Multivariate Analysis, and Bayesian Analysis. MATH324 serves on the order of 120–150 students per year and primarily consists of students majoring in Applied Mathematics and Statistics, Computer Science (in the Data Science track), and Mechanical Engineering (as a technical elective) and, while the dataset and analysis presented is particular to mechanical and materials engineering processes, the physical model is sufficiently straightforward for our student population to be presented without much difficulty.

This article provides an analysis of a large dataset involving the hot-rolling operations of a steel mill located in Decatur,

CONTACT Alexandra M. Newman newman@mines.edu Operations Research with Engineering Graduate Program, Colorado School of Mines, Golden, CO.

Alabama (see supplementary material at the end of the paper for the files). The dataset corresponds to a real-world example of several topics in an introductory regression modeling course taught using texts such as Kutner et al. (2005), Abraham and Ledolter (2006), Montgomery, Peck, and Vining (2012), and Faraway (2015). Topics include: (i) iterative model building to improve goodness-of-fit and prediction, informed by data visualization and model interpretation relative to known physical relationships between model variables; (ii) residual analysis and variable transformations, with considerations of variance stabilizing transformations, the use of underlying physics-based models, and the linearization of multiplicative models; (iii) collinearity and confounding of predictor variables and model interpretation relative to the system being studied, including an interesting illustration of the false correlation paradox; and, (iv) criterion-based variable selection techniques and best-subsets regression.

We present a description of the data and an accounting of the model-building process that is consistent with the university's emphasis on application and interdisciplinary exploration, providing the necessary results and context in an order consistent with the course syllabus and the consultations made with our engineering clients. This article may serve to provide both the framework and a real-word context to introduce these topics in an introductory course on linear regression models. Additionally, the dataset may also lend itself as an introduction to more advanced topics in the statistics curriculum.

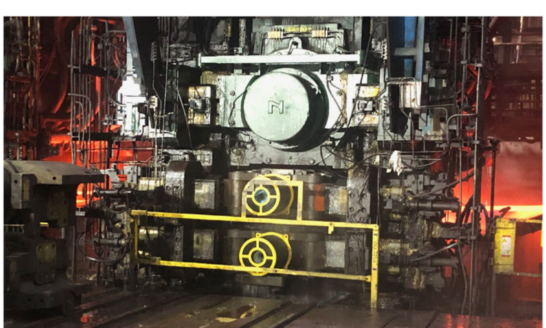
Recent works have employed real-world datasets to present a contextualized construction of statistical models describing phenomena such as marine byozoans (Evans 2022), body fat percentages (Johnson 2021), and Lego prices (Peterson and Ziegler 2021). We add to this literature through the exposition of statistical methodologies employed to develop a regression model to predict the force necessary to properly separate the rolls in the roll stands used to compress relatively thick slabs into sheet steel of a desired thickness (or gauge). This large dataset from an industrial setting represents a unique instructional resource as it provides a "case-study" with which to explore topics found in an introductory regression modeling course. Additionally, this work explores the application of a physics-based model from the manufacturing literature to serve as the basis for a transformation, providing real-world contextualization of important topics in regression modeling (e.g., transformations, variable selection, interpretability).

Because there are variables and factors present in any industrial-scale steel mill that are not accounted for by the available physical models alone, we present a regression modeling approach to predict roll force necessary to achieve the desired gauge drop (or reduction in thickness relative to the coil cast immediately prior in the sequence) for a given coil of steel that depends on coil geometry and steel grade (i.e., chemical composition). From residual plots and knowledge of hot-rolling theory, we introduce a variance stabilizing transformation, consistent with a straightforward physics-based constitutive model of the hot rolling process that includes the same set of variables that are present in the data. From a practical perspective, this approach leads to a predictive model constructed from the data available at the mill that can be interpreted in the context of known physical behavior. Because regression models include both structural and error components, unlike deterministic physics-based equations, predictions based on any of our models can account for uncertainty using the tools of statistical inference.

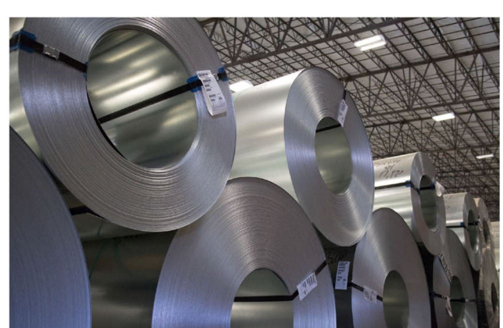
In addition to providing a rich regression modeling example, the confounding of predictor variables in this dataset leads to a powerful illustration of the false correlation paradox, described over a century ago by Yule (1903) and recognized in its simplest form as Simpson's Paradox (Simpson 1951). Our example of this paradox is novel relative to those presented in the existing statistics and data science education literature, for example, Taylor and Mickel (2014), Witmer (2015), and Witmer (2021).

1.1. Overview of Relevant Roll Force Models and Steel Mills

Rolling steel in a hot strip mill is an essential, critically-important process employed in the steel industry (Schey 1983). The required pressure is applied to steel at high temperatures through a series of roll stands in order to reduce the gauge of a steel slab to conform to specifications of a customer order (Figure 1(a)); finished sheets of steel are coiled and delivered to the customer (Figure 1(b)). The pressure, applied as force to each roll stand, is a function of variables such as temperature and steel properties (both chemical and geometric). We present



(a) A steel slab passes through a sequence of roll stands



(b) Finished coils processed for shipping to Nucor customers

Figure 1. Steel hot rolling process. (a) Photo by the author. (b) Photo from the Nucor product website (*https://nucor.com/products/steel/sheet*, accessed November 23, 2023). Used with permission.

an analysis to improve predictions of roll force in an industrial setting and ultimately to better inform operations in a steel mill. Accurate prediction of the relation between roll force and gauge changes as a function of grade and process conditions is essential for safe production– in addition to meeting product quality and thickness requirements. To produce a hot-rolled steel coil, a set of thicknesses and speeds (which affect temperature) in each stand such that the roll force is within the operating limits of the mill stand must be known.

In order to predict the roll force necessary to achieve the desired gauge reduction of a given coil as it passes through a particular roll stand, a model must accurately capture the physics of the hot rolling process and account for sources of variation at industrial scale. There is extensive seminal literature on physics-based computational models of hot rolling steel, for example, Karman (1925), Alexander and Ford (1964), and Alexander (1972). The simple frictionless roll force equation (Kalpakjian and Schmid 2017) considers only basic mechanical and metallurgical properties of the rolled steel, along with geometry of the coil and roll stand.

While they are based on the physics of the process, existing models of roll force are insufficient at the industrial scale because: (i) they neglect some physical inputs for simplicity; (ii) their inputs may not be known or measured; and, (iii) they do not consider potential sources of process variation at the commercial scale. A growing literature on data-informed, empirical models predicts roll force in both hot and cold environments. These models rely on algorithms, heuristics and methodologies that can be challenging to implement and interpret in an industrial-scale environment (Moussaoui, Selaimia, and Abbassi 2006; Jia, Shan, and Niu 2008; Liu, Tong, and Lin 2012; Bagheripoor and Bisadi 2014; Cao et al. 2020). Other models are based on datasets taken from measurements at commercial rolling operations, using regression techniques to extract roll force predictions (Poliak et al. 1998).

Our analysis focuses on the operation at the Nucor Steel facility in Decatur, Alabama. Two electric-arc furnaces melt "heats" of 150–175 tons of a desired grade from scrap metal, which are transferred in ladles to a tundish that feeds two casters that operate in parallel. Each caster produces 98mm-thick slabs with variable width and length according to customer specifications. The slabs are then reheated in parallel tunnel furnaces, which homogenize their internal temperature. Homogenized slabs are transported to a single hot rolling mill, alternating between the two lines. They pass through a roughing mill, where their gauge is reduced to a uniform thickness of approximately 27mm. Each slab then passes through a hot-finishing mill, which consists of five roll stands (Figure 1(a)), each of which contains a pair of rotating rolls, supported by backup rolls, which apply a compressive force. Carefully chosen roll forces incrementally decrease the gauge to its ordered value upon exit from the last roll stand (Stand 5).

The speed of each roll stand must increase to exactly match the increasing speed of the steel strip between each pair of roll stands, according to its decrease in gauge. Any mismatch between the applied roll force and the expected gauge causes a mismatch in velocity, which results in an accumulation or shortage of steel at the next roll stand, producing a sudden catastrophic event called a "cobble," in which steel flies off its prescribed path through the rolling mill. In addition to the expensive shutdown and roll damage, this event constitutes a serious safety hazard; its prevention is contingent upon an accurate model to relate applied roll force and resulting coil gauge under the operating conditions at the time of this writing.

A machine-learning model invoked by the steel mill uses feedback provided by on-line measurement of the gauge at the coiler to inform gauge reductions as each coil passes through the five stands. Maintaining similar forces in each roll stand results in a gauge step-down "policy" across the five roll stands that confounds temperature and strain with roll stand, as we explore in Section 4.2.

After exiting the last stand in the hot rolling mill, the slab (or sheet) is control-cooled with water along a "runout table" to transform the microstructure to achieve desired properties of strength, ductility, and toughness. Finally, it is coiled and either shipped to the customer as a final product, or transported for further downstream processing, such as annealing, coating, and/or cold rolling. Figure 2 illustrates; Allen et al. (2022) and Torres et al. (2024) present a comprehensive description.

1.2. Data Description

Although physics-based models predict the roll force required to achieve the desired gauge reduction at each stand in the five-

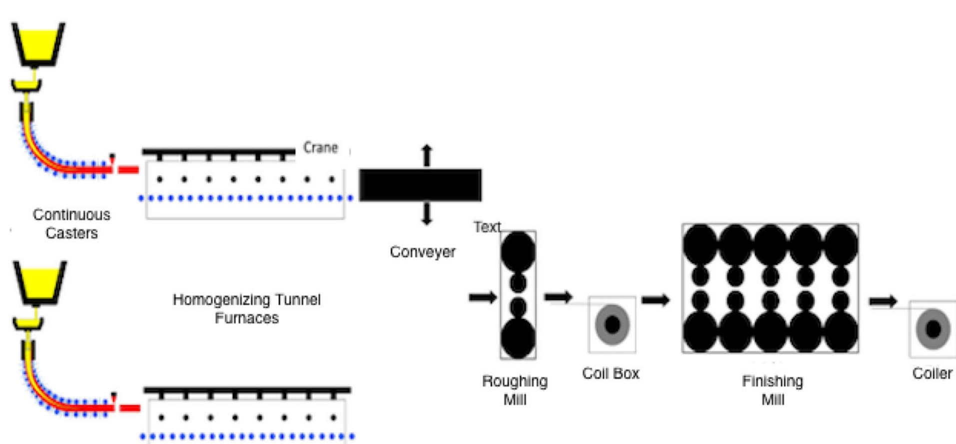


Figure 2. Casting and hot rolling process at the two-caster, single hot-rolling mill (Torres et al. 2024).

Table 1. Variables in the data.

Variable	Categorical or Measured	[Units]
caster identification (North, South)	categorical	
roll stand identification (1, 2, 3, 4, 5)	categorical	
F: force	measured	[N]
w: average slab width	measured	$[mm]$
T: temperature at roll stand	measured	$[K]$
h_0: entry gauge at roll stand	measured	$[mm]$
h_f: exit gauge at roll stand	measured	$[mm]$
r: roll radius $\left(\frac{\text{diameter}}{2}\right)$	measured	$[mm]$

stand hot-finishing-mill, predictions from such models may not match actual measurements taken at the commercial hot-rolling mill. Instead, data from the mill can yield grade-specific regression models to predict force across all five stands for a coil of a given grade (i.e., chemical composition) and geometry (i.e., customer specifications). Force predictions can improve parameter estimates in a mill scheduling optimization model (Torres et al. 2024).

The data provided by the mill, and consistent with existing straight-forward physics-based models for roll force in the steel-making literature, are detailed in Table 1 and include a response variable (force), two categorical predictors (caster location and roll stand location number) and five measured predictors (coil width, temperature, entry gauge, exit gauge, and roll radius), using notation consistent with the steelmaking and hot-rolling literature. This dataset represents a large number of steel coil hot-rolling sequences and includes the following attributes and measurements for 19,960 coils over the five roll stands in the finishing mill for a total of 99,800 data points. In total, the raw data spans 44 grades with the number of coils available for analysis per grade ranging from 13 to 2255. Our work focuses on the grade with the largest number of observations, referred to as Grade 1 in this article and in the linked dataset in the supplementary materials. The dataset for Grade 1 consists of 2255 distinct coils, processed through all five finishing stands in the hot-rolling mill, for a total of 11,275 data points.

We note an opportunity for discussion of independence of observations given that five sets of measurements are recorded on each coil (one set at each of five roll stands through which the gauge of each coil is sequentially reduced), possibly leading to autocorrelation in our model residuals and justifying a consideration of models that incorporate this structure. Because the dataset provided does not include a coil identifier to protect proprietary mill operation details, our analysis excludes a consideration of autocorrelated models. Although not considered in our course, repeated measures models discussed in many texts such as Faraway (2015), Vittinghoff et al. (2012), and Crowder and Hand (1990) might provide an improved model that accounts for both the between-coil and within-coil variability. Within this dataset, each distinct coil has a sequence of five measurements, whose structure is identifiable (e.g., the first five observations represent the first coil as it passes through the five roll stands, the next five observations represent the second coil) to provide the capability of exploring repeated measures models with this dataset in their course.

For the models considered in this article, we note that the material properties of a given coil and the underlying physics of the process are unchanged across the roll stands, suggesting that treating each coil-stand pair as independent is reasonable once we make a variable transformation to incorporate known physical equations into our regression model in Section 3 of this work; the size of our dataset allows for the random partition of the data into training and test sets for model assessment.

Section 2 presents a naïve, additive linear, regression model with only the measured predictor variables **(N1)** and with both the measured predictor variables and the categorical roll stand variable **(N2)**; we introduce model **(T1)** based on a linearization of a physics-based equation to improve model fit and interpretation (Section 3), allowing for a discussion of variable transformations. Unexpected coefficient estimates in **(T1)** lead to a discussion of correlation and the eventual elimination of one predictor in our next model **(T2)** (Section 4.1). Although the models **(T1)** and **(T2)** ignore the categorical roll stand identification variable in an effort to find a model that includes only established physical relationships between force and the measured variables in the data, physically illegitimate coefficient estimates on the temperature term in **(N1)**, **(T1)**, and **(T2)** suggest that roll stand should be considered. This categorical variable is included in our transformed model **(T3)** in Section 4.2, also noting an instance of the false correlation paradox (or Simpson's Paradox) present in **(N1)**, **(T1)**, and **(T2)**. Section 4.3 presents criteria-based variable-selection techniques to justify the elimination of the temperature term in our final model **(T4)**. Section 4.4 discusses model diagnostic plots for **(T4)**. Sections 5 and 6 summarize and conclude, respectively.

2. Initial Naïve Analysis

In developing a linear regression model for roll force, we encounter several concepts through the lens of an interesting, real-world application. Consistent with an early in-class example of multiple regression, we first construct a linear regression model for force as a function of the available predictor variables. We start with pairwise plots of the response variable, force (F), against the measured predictor variables, including width ($X_w = w$), temperature ($X_T = T$), entry gauge ($X_E = h_0$), exit gauge ($X_e = h_f$), roll radius ($X_r = r$), and roll stand (X_S), presented in Figure 3. These graphs show a relationship between force and the measured predictor variables, but with some evidence that the relationships may not be linear and homoscedastic. Moreover, we note the existence of clustering, evident in the plot of force versus temperature, indicating that we need to consider including the categorical stand variable as a predictor. We explore the clustering in this particular graph, along with confounding between the temperature and roll stand variables, in Section 4.2.

Figure 4 plots force against our categorical variables of caster and roll stand in which we see little visual evidence of a meaningful relationship between force and caster, consistent with the underlying physics, so we exclude this variable from consideration. There is strong visual evidence of a relationship between force and roll stand, which we explore in greater detail in Section 4.

Initially, we omit the categorical roll stand variable X_S, and we begin our process with the model that includes only the numeric

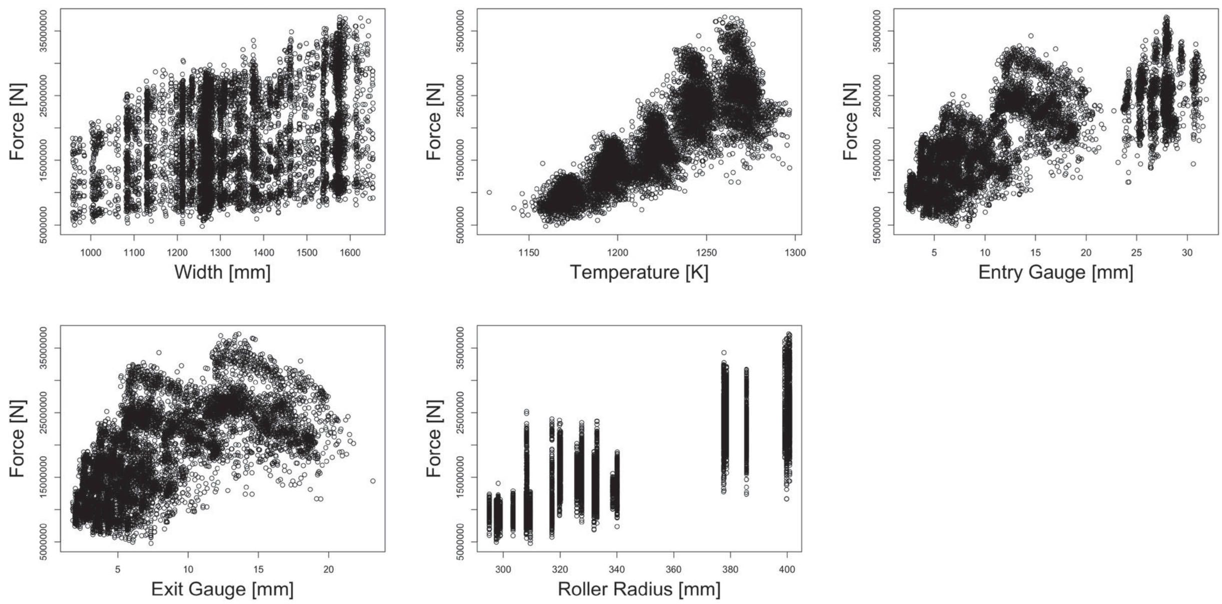


Figure 3. Force versus measured predictor variables.

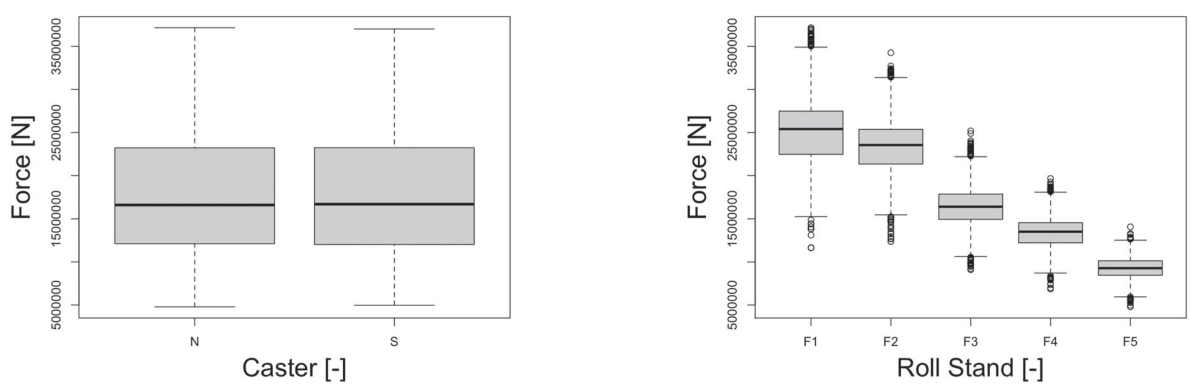


Figure 4. Force versus categorical predictor variables.

Table 2. Regression model summary for **(N1)** ($R^2 = 0.9312$; $F = 30,480$ on 5 and 11,264 degrees of freedom; *p-value* ≈ 0).

Regression coefficient	Estimate	Std. Error	*t statistic*	*p-value*
$\widehat{\beta_0}$	-1.623×10^8	1.321×10^6	-122.82	≈ 0
$\widehat{\beta_w}$	1.471×10^4	1.1116×10^2	131.81	≈ 0
$\widehat{\beta_T}$	1.142×10^5	1.217×10^3	93.82	≈ 0
$\widehat{\beta_E}$	6.828×10^5	1.104×10^4	61.85	≈ 0
$\widehat{\beta_e}$	-1.500×10^6	1.613×10^4	-93.00	≈ 0
$\widehat{\beta_r}$	6.842×10^4	1.253×10^3	54.61	≈ 0

variables. The resulting regression model takes the form:

$$F = \beta_0 + \beta_w X_w + \beta_T X_T + \beta_E X_E + \beta_e X_e + \beta_r X_r + \varepsilon. \quad (1)$$

We refer to (1) as **(N1)**.

Considering the regression output summarized in Table 2, all five predictor variables in **(N1)** are highly significant, with all *t statistics* exceeding 54 in magnitude and *p-values* of essentially zero. Moreover, the overall model is also highly significant, with an *F statistic* of 30,480 and a *p-value* indistinguishable from zero. The impact of sample size on the significance of the model and the parameters cannot be underestimated; the magnitude of the *t statistics* is proportional to the reciprocal of the square root of the sample size. Because this dataset is so large, an absence of significance would arguably be surprising.

While **(N1)** appears to fit well, with $R^2 = 0.9312$, it is imperative to consider whether a better statistical model exists, in terms of goodness-of-fit and prediction statistics, as well as model simplicity and interpretation. The residuals versus fitted values plot in Figure 5 indicates that the standard assumption of homoscedasticity in the errors is likely violated, with potential autocorrelation, suggesting the possible utility of a variance-stabilizing transformation.

Next, we consider a model that incorporates the roll stand in addition to the numeric variables. This regression model takes the form:

$$F = \beta_0 + \beta_w X_w + \beta_T X_T + \beta_E X_E + \beta_e X_e + \beta_r X_r + \sum_{S=2}^{5} \beta_S X_S + \varepsilon. \quad (2)$$

where X_S is a series of indicator variables denoting the stand ($S = 2, 3, 4, 5$) at which the observation is measured. We refer to (2) as **(N2)**.

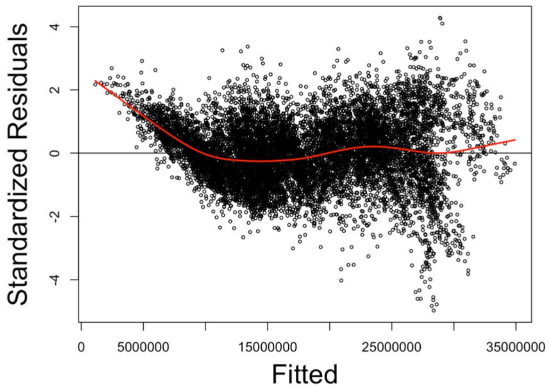


Figure 5. Standardized residuals versus fits plot for **(N1)**.

Table 3. Regression model summary for **(N2)** ($R^2 = 0.9609$; $F = 30,750$ on 9 and 11,260 degrees of freedom; *p-value* ≈ 0).

Regression coefficient	Estimate	Std. Error	*t statistic*	*p-value*
$\widehat{\beta}_0$	4.168×10^7	2.595×10^6	16.06	≈ 0
$\widehat{\beta}_w$	1.494×10^4	8.660×10^1	172.56	≈ 0
$\widehat{\beta}_T$	-2.778×10^4	1.986×10^3	-13.986	≈ 0
$\widehat{\beta}_E$	3.516×10^5	1.415×10^4	24.84	≈ 0
$\widehat{\beta}_e$	-1.2230×10^6	1.439×10^4	-85.01	≈ 0
$\widehat{\beta}_r$	1.948×10^4	2.517×10^3	7.88	≈ 0
$\widehat{\beta}_{Stand2}$	-5.783×10^6	1.321×10^5	-43.79	≈ 0
$\widehat{\beta}_{Stand3}$	-1.336×10^7	2.721×10^5	-49.10	≈ 0
$\widehat{\beta}_{Stand4}$	-1.818×10^7	2.871×10^5	-63.32	≈ 0
$\widehat{\beta}_{Stand5}$	-2.289×10^7	3.602×10^5	-63.56	≈ 0

Considering the regression output summarized in Table 3, we note an improved fit with the inclusion of the roll stand variable in **(N2)**, with $R^2 = 0.9609$. While this highlights the potential impact of clustering by stand, the engineers were primarily concerned about interpretability of the model (specifically, whether the statistical model is aligned with known physical laws), leading us to initially focus on the transformation of variables discussed in Section 3 and then eventually discuss the impact of the stand variable in greater detail in Section 4.2.

We observe that force is measured at each of five roll stands as the gauge of each coil is sequentially reduced, leading us to again note the possibility of constructing a model that accounts for this sequential relation, such as an autoregressive or a repeated measures model. Although these models are not explored here, they are explained and applied to problems elsewhere (Ikram et al. 2016; Huang, Zou, and Song 2021). Here, we note that the material properties of a given coil and the underlying physics of the mechanical deformation behavior are not fundamentally related to the roll stands. This suggests that treating each coil-stand pair as independent is reasonable once we make a variable transformation to incorporate known physical equations into our regression model in Section 3, which aligns with a common variance-stabilizing transformation.

Although our naïve multiple regression model produces coefficients with a straightforward mathematical interpretation, that is, a one-unit change in variable X_i should lead to a β_i unit change in the response variable, the sign on at least one of our model coefficients is in direct contradiction to established knowledge. Specifically, the sign on $\widehat{\beta}_T$ in **(N1)** suggests that, all other things being equal, less force is required to compress steel when it enters an adjacent roll stand with a lower temperature. This unexpected sign difference results from multicollinearity and confounding on the categorical roll stand variable which we discuss in greater detail in Section 4.2. Furthermore, when considering the interpretation of the regression model relative to physical laws associated with the process, we observe that there is no physical model analog in which force has a simple additive linear relationship with the given predictor variables. This suggests a need to consider a transformation of variables that might: (i) accomplish the desired variance stabilization; and, (ii) lead to a statistical model for prediction that incorporates physical properties and is interpretable and consistent with respect to these known properties of the process.

3. Physics-Informed Variable Transformations

Given the concerns with interpretability and heteroscedasticity in the residuals noted in Section 2, we consult a text from Section 1.1 that is accessible to our students (Kalpakjian and Schmid 2017) to identify a physics-based equation, derived from established constitutive models of mechanical behavior during hot rolling of metals such as steel; these models describe the relationship between force and the available predictor variables in a deterministic setting, without the additional unmeasured inputs present at the industrial scale (seen as random variation in our data) that support the use of a statistical model.

3.1. Deterministic Roll Force Equation

Roll force (or roll-separating force) corresponds to the pressure applied by a pair of rolls to flat metal feed in order to reduce the gauge, or thickness, of the coil (see Figure 6). There are several mathematical equations available for calculating roll force during hot rolling, depending on the measurements available (see Section 1.1). Because many of these equations assume knowledge of variable quantities that are not amenable to measurement at the mill (e.g., coefficient of friction), our work uses a simple frictionless roll force equation (Kalpakjian and Schmid 2017), which physically defines the roll-separating force F, as the product of the average flow stress of the material being rolled, $\bar{Y} = \frac{k\epsilon^n}{n+1}$ where $\epsilon = ln\left(\frac{h_0}{h_f}\right)$, w is the width of the metal sheet, and the contact length between the material and the roll is given as $\ell = \sqrt{r(h_0 - h_f)}$ (see Table 4 for variable definitions):

$$F = \bar{Y} w \ell. \tag{3}$$

We note that w and ℓ are entirely associated with the geometry of a given coil, while $\bar{Y}$ is associated with the metallurgical properties of the material, which, in this case, depend on the steel grade and hot-rolling conditions. We use notation consistent with the steelmaking and hot-rolling literature, leading to nonstandard uses of some characters commonly used in the statistics literature. In particular, n represents a parameter associated with the steel's hardening behavior relative to its chemical composition and temperature (as opposed to the sample size associated with data used to construct the statistical model); and, ϵ denotes the strain exerted on the steel which is a function of the reduction in its gauge through a given roll stand, not to

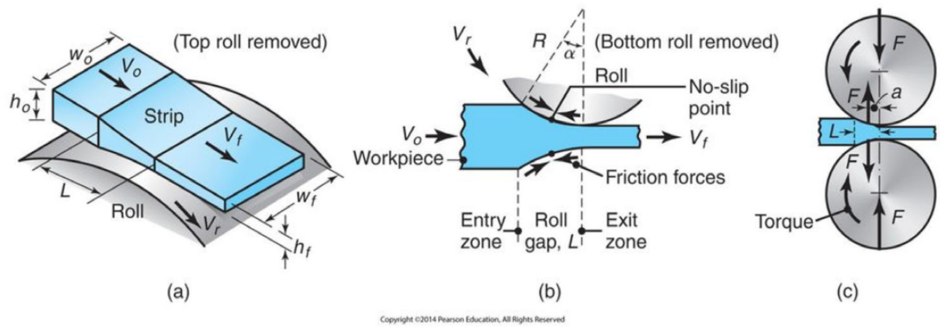


Figure 6. (a) Schematic illustration of the flat-rolling process (b) Friction forces acting on strip surfaces (c) Roll force, F, and torque, τ acting on the rolls (Kalpakjian and Schmid 2017).

Table 4. Roll Force equation notation.

Notation	Definition	[Units]
F	roll-separating force	$[MPa \cdot mm^2 = N]$
h_0, h_f	entry and exit gauges of the strip at the roll stand, respectively	[mm]
$\epsilon = ln\left(\frac{h_0}{h_f}\right)$	strain on coil passing through roll stand	[-]
k	strength coefficient of the steel grade (depends on steel metallurgy and hot-rolling temperature)	[MPa]
n	strain hardening exponent of the steel grade (depends on steel metallurgy and hot-rolling temperature)	[-]
$\bar{Y} = \frac{k\left[\ln\left(\frac{h_0}{h_f}\right)\right]^n}{n+1}$	average flow stress of the material being rolled	[MPa]
w	width of the coil being rolled	[mm]
r	radius of the roll	[mm]
$\ell = \sqrt{r(h_0 - h_f)}$	contact length between the coil and roll	[mm]

Table 5. Strength coefficient notation.

Notation	Definition	[Units]
A	pre-exponential factor (grade dependent)	[MPa]
E_a	activation energy	[MPa·mm^3/moles]
R	universal gas constant (8.314×10^6)	[MPa·mm^3/moles·K]
T	temperature	[K]

be confused with ε, which is standard notation for the random error term in the statistical model.

Although (3) predicts hot-rolling force, it tends to underestimate the actual force in the presence of friction by 15%–20% (Kalpakjian and Schmid 2017) because it fails to account for all contributing variables in the physical system, some of which are impossible to measure in a commercial operation. We do have measurements associated with the geometry of a given coil found in (3) and, although we cannot compute $\bar{Y}$ exactly, we observe that the mechanical behavior of all steel types at hot-rolling temperatures can be generally characterized by an equation of this form, which incorporates the effect of temperature via the following term to express k:

$$k = Ae^{\frac{E_a}{RT}}, \tag{4}$$

where the quantities on the right-hand side of (4) are defined in Table 5.

To arrive at a structural basis for a regression model that relates our response variable to the predictors through a known physical relationship, we substitute (4) into (3) and expand the expression with the quantities in Tables 4 and 5 to obtain the following multiplicative expression of the frictionless roll force model in terms of the predictor variables in our dataset:

$$F = \left(\frac{1}{n+1}\right) \cdot \left(Ae^{\frac{E_a}{RT}}\right) \cdot \left[\ln\left(\frac{h_0}{h_f}\right)\right]^n \cdot w \cdot \ell \tag{5}$$

3.2. Transformation to an Additive Linear Model

Noting that the residuals from our naïve linear regression model in Section 2 suggest a logarithmic variance-stabilizing transformation and that the constitutive model given in (5) is multiplicative on the right-hand side, we construct the strain, $\epsilon = \ln\left(\frac{h_0}{h_f}\right)$, and contact length, $\ell = \sqrt{r(h_0 - h_f)}$, based on our observed variables, as defined in Table 4. (Note that these transformations have a built-in dependency as both are a function of entry and exit gauges of the given rolling stand.) Taking the logarithm of both sides of (5) leads to a linear relationship in the logarithms given by:

$$\begin{aligned} \underbrace{\ln F}_{F^*} &= \underbrace{\ln\left(\frac{A}{n+1}\right)}_{\beta_{0^*}} + \underbrace{\left(\frac{E_a}{R}\right)}_{\beta_{T^*}} \underbrace{\frac{1}{T}}_{X_{T^*}} \\ &+ \underbrace{n}_{\beta_\epsilon} \underbrace{\ln \epsilon}_{X_\epsilon} + \underbrace{1}_{\beta_{w^*}} \underbrace{\ln w}_{X_{w^*}} + \underbrace{1}_{\beta_\ell} \underbrace{\ln \ell}_{X_\ell} \end{aligned} \tag{6}$$

and results in the linear regression model for the transformed variables given by:

$$F^* = \beta_{0^*} + \beta_{T^*}X_{T^*} + \beta_\epsilon X_\epsilon + \beta_{w^*}X_{w^*} + \beta_\ell X_\ell + \varepsilon \tag{7}$$

to which we refer as **(T1)**.

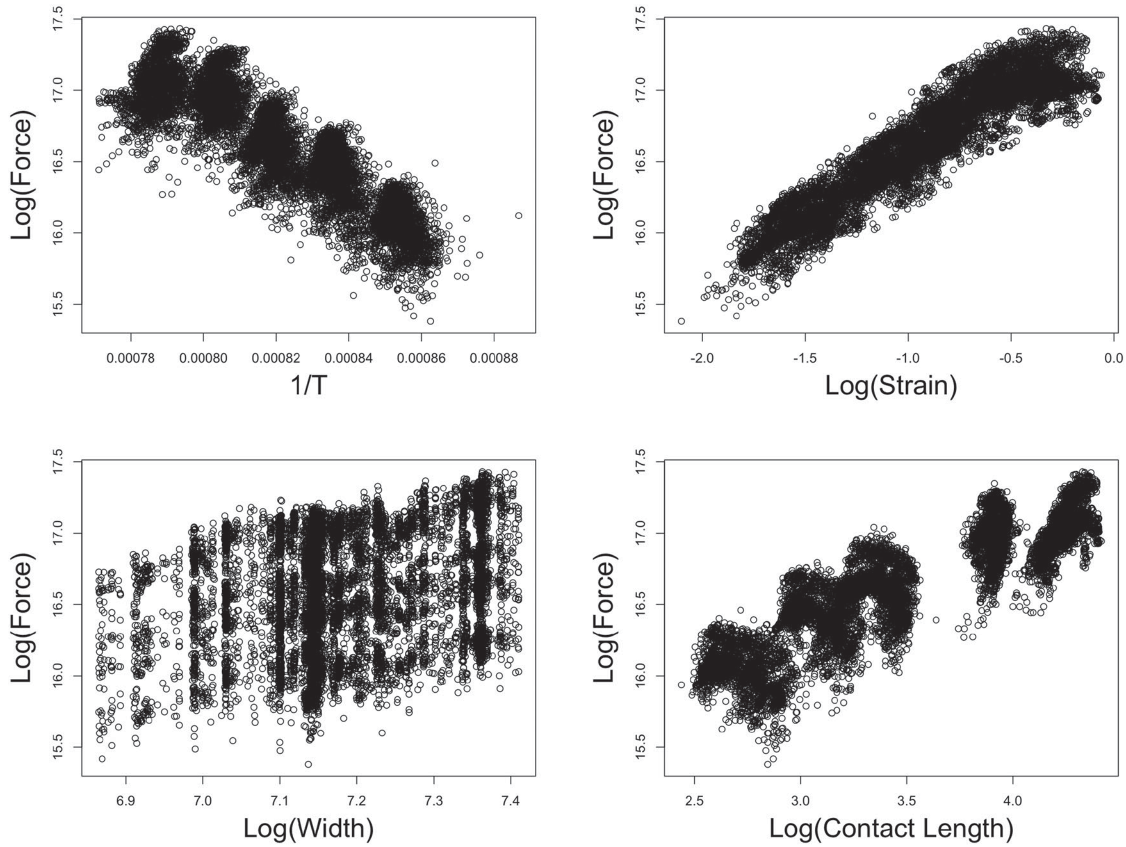


Figure 7. Log force versus log-transformed constitutive model predictor variables.

Although the coefficients in this transformed model do not share the simple "rise over run" interpretation that we associate with the coefficients in **(N1)** and **(N2)**, they are useful and explainable. Structurally, if our data were to conform to the underlying physical model, without any collinearity between variables, we would expect both $\widehat{\beta}_{w^*}$ and $\widehat{\beta}_\ell$ to be close to one. Beyond this, $\widehat{\beta}_{T^*}$ and $\widehat{\beta}_\epsilon$ map onto material properties of the steel being rolled, the activation energy E_a, and the strain-hardening exponent n, respectively. As these quantities vary slightly from heat to heat of the same grade, they are generally unknown at the coil level; this approach to modeling force might produce useful estimates of these quantities for a given coil or heat. This model also presents an opportunity to contextualize variance-stabilizing transformations commonly employed to correct for the violation of homoscedasticity in the errors identified in Figure 5. Observations of an equivalent set of plots of the transformed data (Figure 7) indicate that the log-transformed regression model might improve the fit of the associated statistical model because, to the naked eye, F^* clearly has a more linear relationship, with consistent variability, when plotted against the transformed predictor variables, relative to that observed in Figure 3.

Considering the regression output summarized in Table 6, all four predictor variables in model **(T1)** are significant. We note that the *t statistic* for the coefficient on the contact length term, corresponding to X_ℓ, is 2.681, with *p-value* $= 0.007$ because we would expect model coefficients to have *p-value* ≈ 0 given the size of the sample. Although **(T1)** has one fewer predictor

Table 6. Regression model summary for **(T1)** ($R^2 = 0.9712$; $F = 95,040$ on 4 and 11,265 degrees of freedom; *p-value* ≈ 0).

Regression coefficient	Estimate	Std. Error	*t statistic*	*p-value*	VIF
$\widehat{\beta}_{0^*}$	13.25	0.09896	133.848	≈ 0	
$\widehat{\beta}_{T^*}$	-3931	102.6	−38.314	≈ 0	14.25
$\widehat{\beta}_\epsilon$	0.6761	0.0034	198.374	≈ 0	5.16
$\widehat{\beta}_{w^*}$	1.001	0.0057	175.603	≈ 0	1.03
$\widehat{\beta}_\ell$	0.0113	0.0042	2.681	0.0074	13.40

variables, it fits better to the data, with $R^2 = 0.9712$, than **(N1)** and even **(N2)**. This supports our decision to employ a physics-based model to inform the structure of the statistical model and the appropriate logarithmic transformation. We do note the potential collinearity between predictor variables, evidenced by the variance inflation factor (VIF) values. This collinearity is investigated further in Section 4.1.

The residuals-versus-fits plot for **(T1)** presented in Figure 8 suggests that the variable transformation has improved the underlying assumption of homoscedasticity in the errors relative to **(N1)**. This observation is true of the residual-versus-fits plots for all of the log-transformed models based on the physical equation (5) throughout the rest of this article and is illustrative of: (i) the efficacy of an appropriate variance-stabilizing transformation; and, (ii) the potential to improve model fit using a suitable transformation, specifically, transformations informed by an existing physical model related to the process being considered.

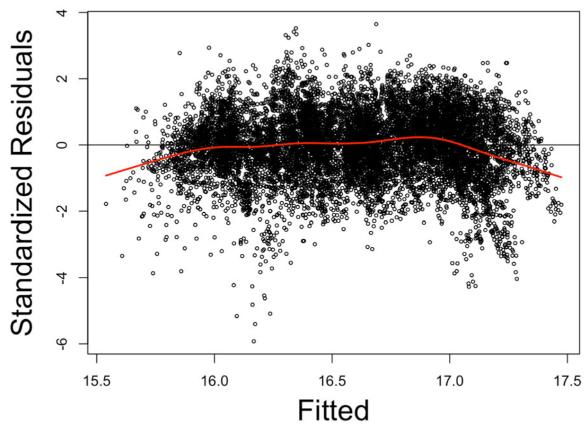


Figure 8. Residuals versus fits plot for **(T1)**.

4. Model Validation and Variable Selection

Having developed a transformation of **(N1)** based on established physical laws, leading to an improved fit in **(T1)**, we turn our attention to ensuring that our statistical model coefficients are consistent with the physics of the hot-rolling mill operations. To the extent that they are not, we investigate the source of any inconsistency. Using the regression model summary output in Table 6, we can determine the estimates of the physical parameters in (5) and (6) (Table 7). Based on (6), we would expect β_{w^*} and β_ℓ both to be approximately equal to 1. Based on physical laws, β_{T^*} should be positive. Based on the material properties of the steel grades processed at this mill, β_ϵ should fall between 0.1 and 10 depending on temperature. As is indicated in bold in Table 7, two of the parameter estimates in **(T1)** are not within the range of values consistent with physical laws. Specifically, $\widehat{\beta_\ell}$ is much smaller than expected, and the sign on $\widehat{\beta_{T^*}}$ is negative, implying that as temperature decreases, less force is required to reduce the gauge of our steel (in direct contradiction to physical laws as well as intuition). This suggests that more investigation is required.

4.1. Collinearity Between Strain and Contact Length

Because the behavior of the coefficient estimates could be explained by multicollinearity, we seek additional improvements in our statistical model. To identify pairwise correlations, we compute the symmetric correlation matrix for the predictor variables used in **(T1)** as

$$\mathbf{C} = \begin{bmatrix} 1.000 & -0.892 & 0.003 & -0.959 \\ & 1.000 & -0.016 & 0.883 \\ & & 1.000 & 0.040 \\ & & & 1.000 \end{bmatrix},$$

where the rows and columns correspond to X_{T^*}, X_ϵ, X_{w^*}, and X_ℓ, in that order. Based on the correlation matrix for these variables, we see that X_{T^*} has a strong negative correlation with both X_ϵ and X_ℓ; X_ϵ has a positive correlation with X_ℓ; and, X_{w^*} does not appear to be correlated with any of the other variables. Another measure of collinearity is the variance inflation factor (VIF) on each of the variables. A strong linear dependence between two or more of our predictors suggests that it may not

Table 7. Regression model coefficient estimates and their physical parameter or mathematical analogues.

Regression coefficient	Computed estimate	Parameter(s) being estimated (and expected range of values)
$\widehat{\beta_{0^*}}$	13.25	$\beta_{0^*} = \frac{A}{n+1} > 0$
$\widehat{\beta_{T^*}}$	**-3931**	$\beta_{T^*} = \frac{E_a}{R} > 0$
$\widehat{\beta_\epsilon}$	0.676	$0.1 < \beta_\epsilon = n < 10$
$\widehat{\beta_{w^*}}$	1.001	$\beta_{w^*} = 1$
$\widehat{\beta_\ell}$	**0.011**	$\beta_\ell = 1$

be necessary to fit a model with all of them, as one or more may be contributing insignificantly. If our predictor variables are linearly independent, then they are orthogonal in the predictor space and the existence of any one variable in the model does not impact the variability of the others. If, however, the predictors are not independent, then the variability of any of our predictors is inflated by the presence of the others, effectively inflating the variance of our estimates.

The VIFs for **(T1)** are given in the rightmost column of Table 6; there is evidence of collinearity, with the largest VIFs of 14.25 and 13.40 corresponding to $\widehat{\beta_{T^*}}$ and $\widehat{\beta_\ell}$, respectively. The VIF may be interpreted as the amount by which the variance of a given estimate is inflated as a result of collinearity. Relative to orthogonal predictors, the VIF on $\widehat{\beta_{T^*}}$ suggests that the standard error on this estimate is $\sqrt{14.25} = 3.77$ times larger that it would have been without collinearity. We also note that the VIF on $\widehat{\beta_{w^*}}$ is approximately one, indicating that it is almost orthogonal to all other predictors, and consistent with its near-zero correlation with the other variables.

Because the influence of temperature and strain on force is determined by established universal physical laws, the engineers prefer to retain these variables in the model. These laws suggest that stress and strain in the processing of steel have a nonlinear dependency on temperature. Hence, large correlations between X_{T^*} and X_ϵ are not surprising. Recalling that $\epsilon = \ln\left(\frac{h_0}{h_f}\right) = \ln(h_0) - \ln(h_f)$ and $\ell = \sqrt{r(h_0 - h_f)}$; that $X_\epsilon = \ln(\epsilon)$; and, that $X_\ell = \ln(\ell)$, we anticipate a pairwise collinearity between these transformed predictors, possibly leading to the overestimation of one and the underestimation of the other. The correlation coefficient of 0.883 in the covariance matrix **C** confirms a fairly strong linear relationship between these two predictors. Finally, if our linear regression model is to agree with the established physical model, then we expect $\widehat{\beta_\ell} \approx 1$. Because $\widehat{\beta_\ell} = 0.011$, this suggests that excluding X_ℓ might produce a model that is in better agreement with the physics. With all of this in mind, we consider a model that excludes X_ℓ in our next iteration, effectively reducing (6) and (7) to:

$$\underbrace{\ln F}_{F^{**}} = \underbrace{\ln\left(\frac{A\cdot\ell}{n+1}\right)}_{\beta_{0^{**}}} + \underbrace{\left(\frac{E_a}{R}\right)}_{\beta_{T^*}} \underbrace{\frac{1}{T}}_{X_{T^*}} + \underbrace{n}_{\beta_\epsilon} \underbrace{\ln\epsilon}_{X_\epsilon} + \underbrace{1}_{\beta_{w^*}} \underbrace{\ln w}_{X_{w^*}} \tag{8}$$

and

$$F^{**} = \beta_{0^{**}} + \beta_{T^*}X_{T^*} + \beta_\epsilon X_\epsilon + \beta_{w^*}X_{w^*} + \varepsilon \tag{9}$$

respectively. We refer to (9) as **(T2)**.

Considering the regression output summarized in Table 8, all three predictor variables in **(T2)** are significant. Interestingly,

Table 8. Regression model summary for **(T2)** ($R^2 = 0.9712$; $F = 126,700$ on 3 and 11,266 degrees of freedom; *p-value* ≈ 0).

Regression coefficient	Estimate	|%| Change from Model (T1)	Std. Error	*t statistic*	*p-value*	VIF
$\widehat{\beta}_{0^{**}}$	13.45	1.5%	0.0624	215.59	≈ 0	
$\widehat{\beta}_{T^*}$	-4154	5.7%	60.21	−68.99	≈ 0	4.90
$\widehat{\beta}_{\epsilon}$	0.6781	0.3%	0.0033	204.09	≈ 0	1.00
$\widehat{\beta}_{w^*}$	1.003	0.2%	0.0056	178.36	≈ 0	4.91

the coefficient estimates for the included variables change very little when X_ℓ is omitted, further justifying the removal of X_ℓ. Although **(T2)** has one fewer predictor variables, there is no loss in goodness-of-fit, with $R^2 = 0.9712$ unchanged from **(T1)**. We also note that the VIFs associated with the parameter estimates have all improved. We consider models that exclude X_ℓ throughout the remainder of our analysis.

4.2. *Confounding of Temperature, Strain, and Roll Stand*

Although we have addressed one of the coefficient estimates that is misaligned with the mathematical expression in our log-transformed physical model, we note that $\widehat{\beta}_\epsilon$ is still larger than material properties would suggest, and that the sign on $\widehat{\beta}_{T^*}$ violates physical laws. The laws of physics dictate that the force required to compress the steel to effect a given gauge reduction should generally increase as temperature decreases (i.e., as $X_{T^*} = 1/T$ increases), suggesting that β_{T^*} should be positive. Because our estimate $\widehat{\beta}_{T^*}$ is significantly less than zero, our model might be improved upon. Reflecting on the graph of force against the ordered categorical roll stand variable in Figure 3, there appears to be a distinct correlation, which is easily confirmed with a simple ANOVA between force and roll stand location. To explore how this may be impacting our estimate of β_{T^*}, we consider a plot of temperature versus roll stand location, given in Figure 9; these variables have a strong relationship, explained by the physics of the environment, as the coil is necessarily cooling as it proceeds from the furnace to the end of the hot-rolling process.

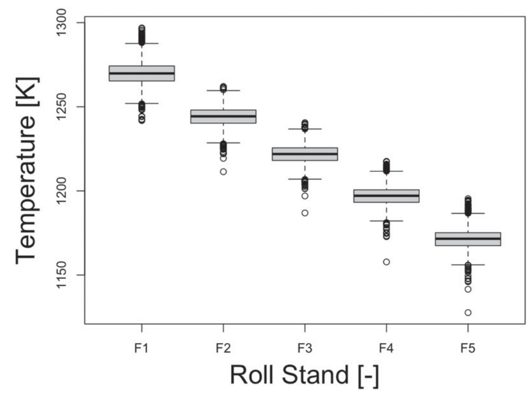


Figure 9. Plot of temperature versus roll stand position.

Based on careful evaluation of the data, leading to discussions with mill engineers, we uncovered a policy reinforced by a proprietary, in-house machine learning model that makes (opaque-to-us) real-time decisions to determine the roll force to apply to each successive coil as it passes through every roll stand. This policy favors larger gauge reductions, which require larger strains and larger exertions of force in the earlier than in the later roll stands to minimize product defects due to excessive wear and, more importantly, malformed steel (cobbles). Figures 10(a) and (b) provide plots of the force and temperature variables used in **(N1)**, **(T1)** and **(T2)**, in which data from the roll stands are assigned a color ranging from hottest (red, stand 1) to coolest (blue, stand 5), representing a standard visualization of a temperature gradient.

The apparent statistical relationship between force and temperature, as well as between the logarithm of force and the reciprocal of temperature, across all stands, possesses a sign opposite of that given by physical laws due to the confounding mill policy of decreasing the roll force with increasing stand (with corresponding decrease in gauge reduction and strain); specifically, the effects of this policy outweigh those of the temperature, which decreases with increasing stand number.

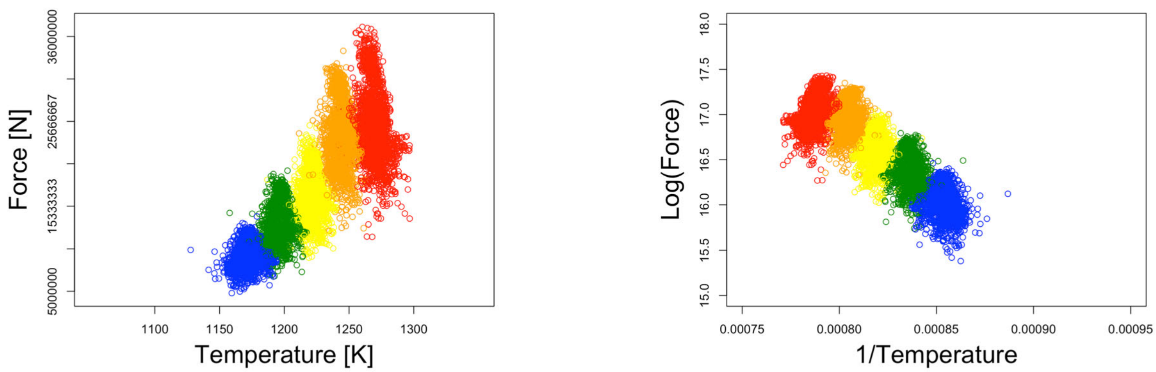
The expected physical relationship of force decreasing with increasing temperature is actually observed at each individual stand, especially at higher temperatures (lower-numbered roll stands) and is consistently opposite the trend seen in the aggregated data over all roll stands (i.e., that force increases with temperature). Figure 10(b) demonstrates more variation in the response than in the predictor across all stands. The combined effect of the relatively narrow range of measured temperatures at each stand and the strong correlation between temperature and strain variables yields mixed results when F^{**} is regressed against X_{T^*} at individual stands, with differing signs on the estimates of $\widehat{\beta}_{T^*}$.

When all stands are combined, however, this same graph indicates a strong negative linear relationship between these variables, consistent with our computed regression coefficient. This is an excellent example (and visualization) of the *false correlation paradox* (Yule 1903). Discussions with engineers suggested that we should include an ordered categorical predictor variable for roll stand, X_S, to account for the stand effect by adding one term to the regression model in (9), leading to a model given by:

$$F^{***} = \beta_{0^{***}} + \beta_{T^*}X_{T^*} + \beta_\epsilon X_\epsilon + \beta_{w^*}X_{w^*} + \sum_{S=2}^{5} \beta_S X_S + \varepsilon \quad (10)$$

where X_S is a series of indicator variables denoting the stand $(S = 2, 3, 4, 5)$ at which the observation is measured, with temperature decreasing monotonically as the stand indicator increases for a given coil. We refer to (10) as **(T3)**.

Output from **(T3)** summarized in Table 9 shows a modest improvement in fit, with $R^2 = 0.9761$. Having added parameters in **(T3)**, we note that $R^2_{adj} = 0.9761$, indicating that the addition of roll stand does not, in and of itself, contribute to overfitting



(a) Force versus Temperature (b) Log(force) versus 1/Temperature

Figure 10. Plots of force and temperature variables colored by roll stand position.

Table 9. Regression model summary for **(T3)** ($R^2 = 0.9761$; $F = 65{,}660$ on 7 and 11,262 degrees of freedom; *p-value* ≈ 0).

Regression coefficient	Estimate	Std. Error	*t statistic*	*p-value*	VIF
$\widehat{\beta}_{0^{***}}$	10.137	0.1121	90.439	≈ 0	
$\widehat{\beta}_{T^*}$	**79.126**	**135.300**	**0.585**	**0.559**	29.81
$\widehat{\beta}_{\epsilon}$	0.604	0.0037	194.069	≈ 0	7.22
$\widehat{\beta}_{w^*}$	0.996	0.0051	194.069	≈ 0	1.00
$\widehat{\beta}_{Stand2}$	−0.060	0.0029	−20.719	≈ 0	3.90
$\widehat{\beta}_{Stand3}$	−0.186	0.0048	−38.609	≈ 0	10.85
$\widehat{\beta}_{Stand4}$	−0.231	0.0072	−32.109	≈ 0	24.11
$\widehat{\beta}_{Stand5}$	−0.376	0.0099	−37.976	≈ 0	45.78

in our model. With the introduction of the categorical roll stand variable, we also note that the intercept term, $\beta_{0^{***}}$, now includes the Stand 1 effect, while the effects for Stands 2–5 are reported in the summary regression output in Table 9. Interestingly, when roll stand is included, we see that the coefficient estimate on the temperature variable now has the correct sign ($\widehat{\beta}_{T^*} = 79.126 > 0$), but it is no longer significantly different from zero (*p-value* $= 0.559$); the roll stand location effect dominates the temperature effect, which also explains the negative coefficient on the X_T variable in the models that omitted the stand variable. Finally, we note that the VIFs for this model are quite large, consistent with the confounding of temperature and roll stand previously discussed. Given the *p-value* on $\widehat{\beta}_{T^*}$ and the large VIFs, we explore variable selection techniques to determine whether any additional variable eliminations are warranted.

4.3. Variable Selection Techniques to Eliminate the Temperature Term

The temperature term becomes insignificant with stand explicitly included in the model. Correspondingly, variable selection techniques can employ criterion-based procedures. We use the R^2_{adj} and Mallow's C_p statistics to determine the number of predictors to include, and *best subsets* to determine the best model with a given number of predictors. Figures 11(a) and 11(b) present plots of these criteria against the number of predictors in the best fitting model of each size.

Because adding a variable to the model only increases the R^2_{adj} statistic when that variable has additional explanatory value, we choose model variables that maximize this criterion. Based on the plot in Figure 11(a), the maximum value of R^2_{adj} appears to occur with six or seven variables. Using the "which.max()" function in **R**, we learn that the maximum value occurs in the six-predictor model, that is, the model with one of the predictor variables removed from **(T3)**.

For the purpose of comparison, we consider Mallow's C_p as an alternative criterion to determine the best model size. Figure 11(b) provides a plot of this statistic against the number of predictor variables (p) in the model, as well as the line $C_p = p$. Using this criterion, we select the first model for which $C_p \le p$ (i.e., the value of p for which the C_p statistic is on or below the line). We see that this occurs for the six-predictor model, confirming the results using the R^2_{adj} as our criterion.

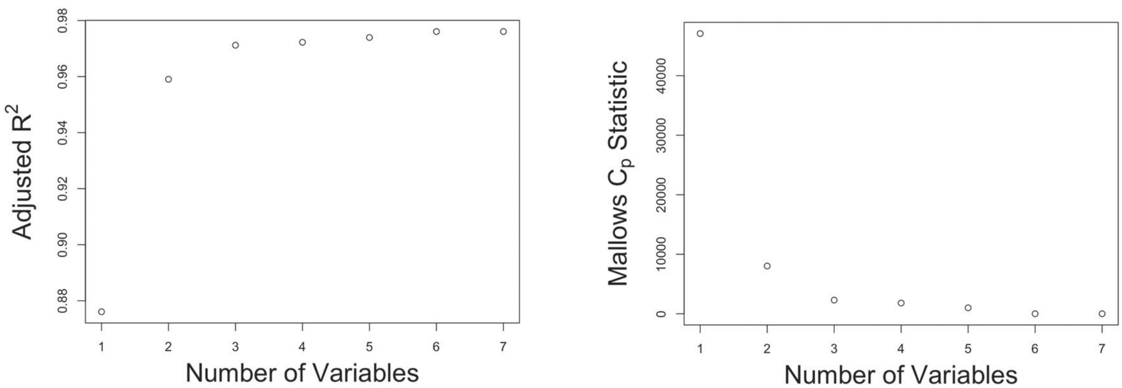
Both criterion-based selection techniques indicate that the six-predictor model substitutes reasonably for **(T3)**. Table 10 summarizes the best subsets regression, using **(T3)** as the full model and the R^2_{adj} statistic for our measure of model quality. This table indicates which variables are included in the best fitting model of each size with a checkmark in the column associated with that variable. (The Stand 1 indicator variable, X_{S_1}, is included in the intercept term, which is present in all of these models.) The highlighted row in our table corresponds to the six-predictor model with the largest R^2_{adj} statistic, which possesses all predictor variables except temperature. The criterion-based approach suggests the removal of the temperature term from **(T3)**.

Because we are interested in making predictions with this model, statistics such as the PRESS or R^2_{pred} could also be used as criteria for variable selection. The slight improvement in the R^2_{pred} statistic for **(T4)** over **(T3)** in Table 12 in Section 4.5 suggests that we would arrive at the same model with these other statistics.

Our final iteration includes X_S to account for the stand effect, but eliminates the temperature term X_{T^*} from the regression model in (10), leading to the model given by:

$$F^{****} = \beta_{0^{****}} + \beta_{\epsilon} X_{\epsilon} + \beta_{w^*} X_{w^*} + \sum_{S=2}^{5} \beta_S X_S + \varepsilon \qquad (11)$$

to which we refer as **(T4)**.



(a) R^2_{adj} versus number of predictors

(b) C_p versus number of predictors

Figure 11. Plots of variable selection criteria.

Table 10. Best model of each size based on the R^2_{adj} statistic.

Number of Predictors (p)	Temp. (X_{T^*})	Strain (X_ϵ)	Width (X_{w^*})	Stand 2 (X_{S_2})	Stand 3 (X_{S_3})	Stand 4 (X_{S_4})	Stand 5 (X_{S_5})
1		✓					
2		✓	✓				
3	✓	✓	✓				
4	✓	✓	✓			✓	
5		✓	✓		✓	✓	✓
6		✓	✓	✓	✓	✓	✓
7	✓	✓	✓	✓	✓	✓	✓

Table 11. Regression model summary for **(T4)** ($R^2 = 0.9761$; $F = 76,610$ on 6 and 11,263 degrees of freedom; *p-value* ≈ 0).

Regression coefficient	Estimate	|%| Change from Model (T3)	Std. Error	*t statistic*	*p-value*	VIF
$\widehat{\beta}_{0^{****}}$	10.198	0.6%	0.0368	277.03	≈ 0	
$\widehat{\beta}_\epsilon$	0.604	0.0%	0.0037	164.36	≈ 0	7.22
$\widehat{\beta}_{w^*}$	0.996	0.0%	0.0056	194.12	≈ 0	1.00
$\widehat{\beta}_{Stand2}$	−0.059	1.7%	0.0019	−31.61	≈ 0	1.60
$\widehat{\beta}_{Stand3}$	−0.184	1.1%	0.0024	−77.02	≈ 0	2.65
$\widehat{\beta}_{Stand4}$	−0.227	1.7%	0.0031	−73.39	≈ 0	4.46
$\widehat{\beta}_{Stand5}$	−0.371	1.3%	0.0043	−87.08	≈ 0	8.47

Based on the summary of **(T4)** in Table 11, we note that the $R^2 = 0.9761$ statistic for this model is unchanged relative to **(T3)**, which includes the temperature term. Although **(T4)** possesses two of the variables related to the measured quantities in our dataset, inclusion of the categorical roll stand variable produces a model with the best fit among all considered based on the R^2 goodness-of-fit measure. All of the regression coefficients are highly significant, as is the overall model. We also note little change in the parameter estimates for the variables included in both **(T3)** and **(T4)**, confirming that the temperature contributes very little to the model when the categorical roll stand variable is included. And, there is a marked reduction in the VIFs across all of the variables included relative to **(T3)**. Additionally, the inclusion of the categorical roll stand variable, and resulting exclusion of the temperature variable, is one way to resolve the most troubling inconsistency between our regression model and physical laws by effectively incorporating mill policies into our model. Another way to resolve this inconsistency could be to further improve the equation form by calculating and including the variable of strain rate, which also varies greatly between roll stands and is known to be important to the mechanical property behavior. The associated advanced transformations are beyond the scope of the linear regression models covered in our introductory course, so this is suggested as future work.

4.4. Model Diagnostic Plots

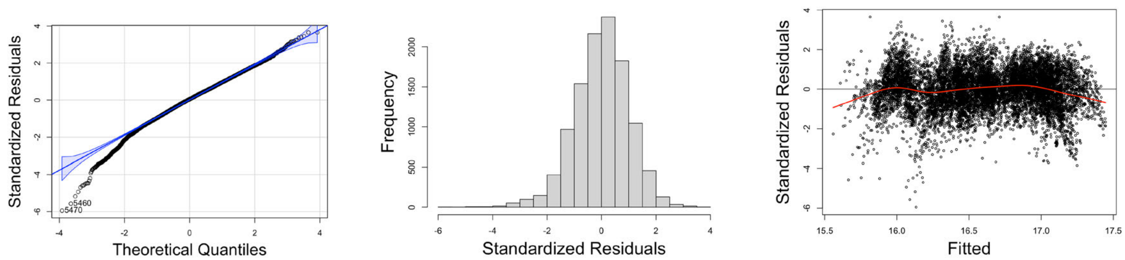
Finally, we present residual plots to assess the standard assumptions about the errors in model **(T4)**, (i) ε_i are independent, (ii) $\varepsilon_i \sim N(0, \sigma^2)$ and (iii) ε_i are homoscedastic. Figures 12(a)–(c) present the Normal Q-Q plot, the histogram and the residuals versus fits plot for this model, respectively.

The assumption of normality is questionable because the model produces negative residuals that are larger in magnitude (in their most extreme cases) than is seen in the positive residuals, resulting in: (i) a Q–Q plot that is not linear; and, (ii) a histogram that is not symmetric about zero. If the model is used to make inferences, resampling techniques should be considered as an alternative to formulas that rely on classical normal distribution theory. We also note the large number of negative outliers in the Q-Q plot. Based on our discussions with plant engineers, it appears that the machine learning model favors very low strains (and, consequently, forces) in the last of the five roll stands to prevent defects in the finished coils, likely resulting in overestimates of force at this stage at this stand. This calls for an investigation of unexpected behaviors in the residuals.

The residuals versus fitted values plot in Figure 12(c) shows no discernible pattern, demonstrates a fairly even spread across the horizontal axis, and supports the assumptions of independence and homoscedasticity in the errors. Because the equivalent plot for **(N1)** demonstrates heteroscedasticity, our variable transformations mitigate this characteristic and lead to a better fitting model.

4.5. Model Quality for Prediction

Given the first-order objective of developing a regression model to better predict roll force across the roll stands in the mill given



(a) Normal Q-Q plot (b) Histogram (c) Residuals versus fits plot

Figure 12. **(T4)** standardized residuals plots.

Table 12. Models considered with variables included, R^2 and R^2_{pred} statistics.

Model	Variable Included					R^2	R^2_{pred}
	X_{T*}	X_ϵ	X_{W*}	X_ℓ	X_S		
(N1)	All untransformed measured variables					0.9312	0.931073
(N2)	All untransformed variables					0.9609	0.960816
(T1)	✓	✓	✓	✓		0.9712	0.971193
(T2)	✓	✓	✓			0.9712	0.971180
(T3)	✓	✓	✓		✓	0.9761	0.976045
(T4)		✓	✓		✓	0.9761	0.976050

the available predictor variables, we conclude by presenting an assessment of the prediction quality of the five models developed in this article. A standard textbook approach computes the Prediction Error Sum of Squares (PRESS) which is the sum of the squared differences between the ith observation of the response in our data (F_i) and the predicted value of that response using the model fit without the ith case ($\widehat{F}_i$).

$$\text{PRESS} = \sum_i \left(F_i - \widehat{F}_i\right)^2.$$

However, because one of our models is on the original scale of force and the others are on a logarithmic scale, the dependence of the PRESS statistic on units is problematic, so we instead use the unit-free prediction R^2, which is computed as

$$R^2_{pred} = 1 - \frac{PRESS}{SST},$$

where $SST = \sum_i \left(F_i - \bar{F}\right)^2$ and $\bar{F}$ represents the mean value of the sample. Table 12 summarizes the models introduced in this article, including the variables in each model, R^2, and R^2_{pred}.

The R^2 statistic is consistently high and increases as we progress through our iterative approach to model selection. A similar trend is present in the R^2_{pred} statistic, which is a measure of how well our model predicts the response variable associated with roll force. We note that R^2_{pred} generally increases throughout the model development presented in this article and the largest observed value corresponds to our final model, **(T4)**, which has the added advantage of consistency with physical laws. Based on the values of both R^2_{pred} and R^2 across all five models, we are confident of an absence of overfitting.

To further assess the predictive quality of this statistical modeling approach, we randomly split the data into training and testing sets, consisting of 80% and 20% of the data (1804 and 451 coils), respectively. **(T4)** was reconstructed using the training data and employed this model to predict the log($force$) response variable in the test data. Figure 13 presents a plot of predicted vs. observed $log(force)$ using **(T4)** for this test data. This plot confirms that **(T4)** is a quality predictive model, supported by an $R^2 = 0.9777$ in the linear fit between predicted and observed values in the test dataset.

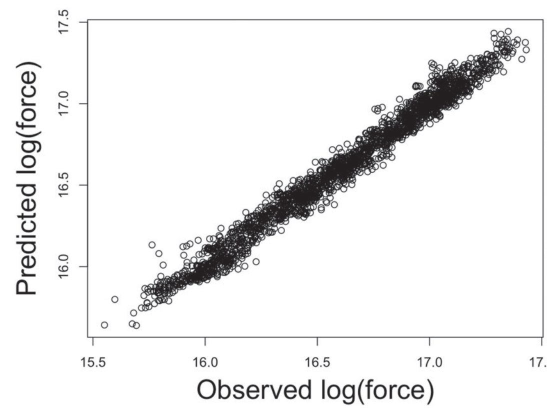


Figure 13. Plot of predicted versus observed log(force) on the test dataset.

Finally, to provide a more realistic comparison for the value of the physics-based transformation, we can consider a variant of **(N2)** in which the left-hand side is log-transformed but the right-hand side is still additive and includes variables consistent with **(T4)**:

$$\ln F = \beta_0 + \beta_\epsilon X_\epsilon + \beta_W X_w + \sum_{S=2}^{5} \beta_S X_S + \varepsilon. \tag{12}$$

Absent the presence of the frictionless roll force model to motivate our transformation, this model is consistent with what would be discussed to stabilize variance and incorporate the categorical roll stand variable in many introductory regression courses. We note that $R^2 = R^2_{pred} = 0.9606$ for this model, indicating that **(T4)** is still the better model.

5. Overview of Iterative Model Building

While this dataset certainly provides opportunities to explore additional topics, our scope here has been the development and analysis of linear regression models for the purpose of predicting the roll force required to accomplish a desired gauge reduction for a steel coil with a given geometry and chemical composition. A secondary goal is to create a model that is reasonably consistent with the physical laws governing the hot-rolling process. To accomplish these objectives, we explore a variety of topics of interest in the instruction of a course using a practical application.

Toward the end of achieving the objective of identifying a good regression model for the prediction of roll force in the steel mill, we began with naïve linear regression approaches in Section 2, using the measured variables in our dataset as the predictors in **(N1)** and the measured variables plus the categorical roll stand variable in **(N2)**. Although these models have strong goodness-of-fit characteristics ($R^2 = 0.9312$ and $R^2 = 0.9609$, respectively), they do not adequately align with the governing physical principles and exhibit heteroscedasticity in their residual plots.

Section 3 demonstrates that it is good practice to employ known physical laws in the construction of a model (when possible), and employs a simple equation based on rolling theory from a manufacturing engineering text involving the measured variables in our dataset. This model requires a relatively straightforward logarithmic variable transformation that also serves as variance-stabilizing, addressing our concerns about homoscedasticity of the model errors. Additionally, this transformed model, **(T1)**, shows improvement in goodness-of-fit ($R^2 = 0.9712$). However, some fitted parameters are inconsistent with the underlying deterministic physical model.

Section 4 considers a variety of approaches to validate our model against the underlying physics, leading to: (i) the elimination of the contact length variable in **(T2)** as that variable was highly correlated with two other variables, and was mathematically "similar" to the strain variable in its construction; (ii) the subsequent inclusion of a categorical variable for roll stand in **(T3)** as roll stand is strongly correlated with temperature and provides a good explanation of the seeming violation of physical laws; and, (iii) the ultimate exclusion of the temperature variable in **(T4)** based on standard variable selection techniques and supported by the interaction of necessary monotonic temperature drops across the roll stands and based on mill policy. We note that **(T4)** demonstrates exceptional goodness-of-fit ($R^2 = 0.9761$) and is far more consistent with the underlying deterministic physical model than the other models presented. Further model improvements could be made using tools that are beyond our course.

6. Concluding Remarks

The story associated with the analysis of this dataset presents a contextualized example of an iterative approach to regression modeling informed by a number of methodologies common to an applied course. This large dataset from an industrial setting is rich in that its naïve analysis points to the necessity of considerations beyond simply fitting a response variable to the available predictors in a strictly linear form.

The development of an accurate model of roll force is of practical importance to industry. In addition to controlling the operation to avoid cobbles, the rolls also wear in proportion to the roll force. Each slab deforms the roll across the contact area, which causes profile degradation of the roll in proportion to the applied force. Consequently, each roll should be replaced before it causes defects in the product. An accurate model of roll force can enable accurate prediction of roll lifetime, with significant benefits in product quality, mill scheduling, production, and maintenance. Because our original objective was to construct a regression model for roll force prediction during hot rolling of steel, the size of this dataset is conducive to separating it into a test dataset and training our models on a randomly selected subset (Picard and Cook 1984).

In consultation with engineering colleagues and literature on mechanistic models of hot steel rolling, physically-based models provide a structure for formulation. Whereas many require inputs not explicitly available in the dataset and some introduce a mathematical structure that cannot be transformed into a linear regression form, we restrict ourselves to a frictionless strain-based force model. While physical models are generally elegant, the statistical model, through its error term, is able to account for variables that are not present in these physical models, leading to the tension between the models discussed in this article and the underlying physical model that motivated the transformation presented. We concede that there are other, more advanced physically-based models which could be explored in future work, for example, nonlinear or autocorrelated.

In the semester, current at the time of this writing, the instructor of our introductory course developed assignments through which the students revisit the dataset as a real-world case study throughout the term. To date, the students have completed two assignments based on this dataset. Prior to the initial assignment, the dataset was shared with the students and a portion of class time was devoted to a description of the data, including definition of variables and a description of the process from which it came. This introduction was supported by Section 1 of this article. The instructor leveraged Figures 3 and 6 to describe the variables in the dataset and the students indicated comfort with these descriptions. Additionally, because this dataset is being used as a case study in support of better understanding topics as they are introduced, assignments have been created as **R** notebooks and are presented and discussed in class when they are assigned and when they are graded, revisiting the dataset and associated variable definitions, as well as work done to date on this data. The two initial assignments are included in pdf form in the supplementary materials.

The initial assignment, titled "Steel Activity 1," is a simple exploratory data analysis in which students generate pairwise plots of the data, similar to those found in Section 2, and then construct two simple linear regression models, consistent with what they were currently studying in the course. Based on their plots, students are asked to choose one quantitative variable to predict force and, separately, one categorical variable to predict force, including justifications for their choices. They are then asked to construct two models, one to predict force based on a single quantitative variable (Temperature) and the other to predict force based on a single categorical variable (Stand). In each instance, they are asked to respond to open-ended questions regarding model fit, validity and interpretation. Responses include concerns about the plots not being very linear in a few cases, with some students identifying heteroscedasticity. A few students correctly note that the plot between force and temperature seems to have the wrong slope based on their understanding of the influence of temperature on physical material properties.

Based on their responses to these questions, the instructor is able to lead a contextualized discussion that addresses the equal variance assumption and interpretability of statistical models. Moreover, based on the discussion of confounding between temperature and stand discussed in Section 4.2, the instructor is able to begin a conversation about confounding of variables and lurking variables.

The second assignment, titled "Steel Activity 2," is an extension of " Steel Activity 1" that asks the students to consider multiple regression models (without transformation). In this assignment, they are asked to fit a full first-order model with all of the quantitative variables, similar to **(N1)** in Section 2. They are asked to discuss whether model conditions appear to be met and whether they have concerns about multicollinearity, including a computation and discussion of variance inflation factors. Finally, they are asked to create a multiple regression model that includes the quantitative variables plus the categorical Stand variable and to interpret the coefficient on temperature within the context of this model, as well as what they understand to be the relationship between the force required to deform a material and the temperature of that material, again anticipating the discussion found in Section 4.2. Based on this assignment, the instructor is able to lead a contextualized discussion of violations of model assumptions, multicollinearity and interpretability of model coefficients, all in anticipation of a final assignment that leverages Sections 3 and 4 to (i) identify a model with interpretable coefficients that also conforms to a standard variance stabilizing transformation, (ii) eliminate variables to address multicollinearity, and (iii) provide a good example of a variable selection process that identifies a regression model that balances model size and fit based on multiple criteria and identifies a model whose interpretation is sensible in the context of physical laws and a real-world dataset.

Given the composition of our student population, the dataset has been received well and the instructor has had robust classroom discussions about data, variable definitions, potential models and their challenges, all contextualized by this new dataset from an industrial partner affiliated with the school. Based on initial feedback and class discussion, our students' past experience with interdisciplinary STEM projects, and the emphasis on modeling and contextualization throughout our core curriculum, the physical model presented in Section 3 and the associated transformation and variable selection discussion presented in Section 4 anchor the topics in our introductory regression course in a meaningful way.

Supplementary Materials

Hot-Roll Mill Roll Force Dataset: Gr01Data.csv
This dataset is real-world data from the Nucor Decatur, Alabama facility. Identifier variables referencing proprietary information have been removed. This is the complete dataset used for the analysis presented in this article.
R code used for the analysis in this article: Grade01Analysis.R
Codebook for variables and models described in this article: Gr01_Codebook.xlsx
Associated Assignments: Steel Activity 1 and Steel Activity 2

Acknowledgments

The authors wish to thank their industry contacts at Nucor Steel, Decatur, Alabama, USA, in particular, Dr. Bryan Petrus, for their help with and insights on all industry-specific aspects of this project. We acknowledge the aid of Dr. Ashlyn Munson and Dr. Daisy Philtron for their feedback on the presentation of the iterative regression modeling approach as it relates to their in-class experiences at the Colorado School of Mines and their collaboration in designing student activities around this dataset. Permission to make the dataset available (in its redacted form) was kindly granted by *Nucor Steel, Decatur Alabama* and the *Continuous Casting Center at the Colorado School of Mines* directed by Professor Brian G. Thomas.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

This work was partially funded by the Continuous Casting Center at the Colorado School of Mines and by ANID/CONICYT-Fondecyt Regular 1200809.

Data Availability Statement

Data available within the article or its supplementary materials: The authors confirm that the data supporting the findings of this study are available within the article and/or its supplementary materials.

References

Abraham, B., and Ledolter, J. (2006), *Introduction to Regression Modeling*, Pacific Grove, CA: Thomson Brooks/Cole.
Alexander, J. M. (1972), "On the Theory of Rolling," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 326, 535–563.
Alexander, J. M., and Ford, H. (1964), "Simplified Hot-Rolling Calculations," *Journal of the Institute of Metals*, 92, 397–404.
Allen, M., Tarvin, D. A., Newman, A. M., and Thomas, B. G. (2022), "Nucor Steel Improves Production Scheduling, working paper.
Bagheripoor, M., and Bisadi, H. (2014), "An Investigation on the Roll Force and Torque Fluctuations During Hot Strip Rolling Process," *Production and Manufacturing Research*. 2, 128–141.
Cao, X., Zhu, G., Yue, B., Qiao, S., Gao, X., Guo, N., and Xu, M. (2020), "Research on Cast-Rolling Force in Twin-Roll Continuous Strip Casting Based on Uncertainty Analysis," *Materials Today Communications*, 24, 101100.
Crowder, M., and Hand, D. (1990), *Analysis of Repeated Measures* (1st ed.), New York: Routledge.
Evans, C. (2022), "Regression, Transformations, and Mixed-Effects with Marine Bryozoans," *Journal of Statistics and Data Science Education*, 30, 198–206.
Faraway, J. J. (2015), *Linear Models with R* (2nd ed.), New York: Taylor and Francis.
Huang, J.-B., Zou, H., and Song, Y. (2021), "Biased Technical Change and its Influencing Factors of Iron and Steel Industry: Evidence from Provincial Panel Data in China," *Journal of Cleaner Production*, 283, 124558.
Ikram, A., Su, Q., Rafiq, M. Y., and Ramiz-Ur-Rehman (2016), "Time Series Modelling for Steel Production," *The Journal of Developing Areas*, 50, 191–207.
Jia, C.-y., Shan, X.-y., and Niu, Z.-p. (2008), "High Precision Prediction of Rolling Force Based on Fuzzy and Nerve Method for Cold Tandem Mill," *Journal of Iron and Steel Research International*, 15, 23–27.
Johnson, R. W. (2021), "Fitting Percentages of Body Fat to Simple Body Measurements: College Women," *Journal of Statistics and Data Science Education*, 29, 304–316.
Kalpakjian, S., and Schmid, S. (2017), *Manufacturing Processes for Engineering Materials* (6th ed.), Harlow: Pearson.
Karman, T. V. (1925), "On the Theory of Rolling," *A Angewandte Mathematik und Mechanik*, 5, 130–141.

Kutner, M. H., Nachtcheim, C. J., Neter, B., and Li, W. (2005), *Applied Linear Statistical Models* (5th ed.), Irwin, NY: McGraw Hill.

Liu, Y., Tong, C., and Lin, F. (2012), "Rolling Force Prediction Based on Wavelet Multiple-RBF Neural Network," in *2012 IEEE International Conference on Information and Automation*, pp. 941–944.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012), *Introduction to Linear Regression Analysis* (5th ed.), Hoboken, NJ: Brooks/Cole.

Moussaoui, A., Selaimia, Y., and Abbassi, H. A. (2006), "Hybrid Hot Strip Rolling Force Prediction Using Bayesian Trained Artificial Neural Network and Analytical Models," *American Journal of Applied Sciences*, 3, 1885–1889.

Peterson, A. D., and Ziegler, L. (2021), "Building a Multiple Linear Regression Model with Lego Brick Data," *Journal of Statistics and Data Science Education* 29, 297–303.

Picard, R. R., and Cook, R. D. (1984), "Cross-Validation of Regression Models," *Journal of the American Statistical Association*, 79, 575–583.

Poliak, E. I., Shim, M. K., Kim, G., and Choo, W. (1998), "Application of Linear Regression Analysis in Accuracy Assessment of Rolling Force Calculations," *Metals and Materials*, 4, 1047–1056.

Schey, J. A. (1983), *Tribology in Metalworking: Friction, Lubrication and Wear*, Metals Park, OH: American Society for Metals.

Simpson, E. H. (1951), "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society*, Series B, 13, 238–241.

Taylor, S. A., and Mickel, A. E. (2014), "Simpson's Paradox: A Data Set and Discrimination Case Study Exercise," *Journal of Statistics Education*, 22, 1–18.

Torres, N., Greivel, G., Betz, J., Moreno, E., Newman, A., and Thomas, B. (2024), "Optimizing Steel Coil Production Schedules Under Continuous Casting and Hot Rolling," *European Journal of Operational Research*, 314, 496–508.

Vittinghoff, E., Glidden, D., Shiboski, S., and McCulloch, C. (2012), *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*, Statistics for Biology and Health, New York: Springer. *https://books.google.com/books?id=boMUM6N8V4wC*

Witmer, J. (2015), "How Much Do Minority Lives Matter?" *Journal of Statistics Education*, 23, 1–5.

——— (2021), "Simpson's Paradox, Visual Displays, and Causal Diagrams," *The American Mathematical Monthly*, 128, 598–610.

Yule, G. U. (1903), "Notes on the Theory of Association of Attributes in Statistics," *Biometrika*, 2, 121–134.