VARIABLE SELECTION USING GAUSSIAN COPULA REGRESSION

by

Lucas Keetch

Submitted in partial fulfillment of the requirements
for the degree of Bachelor of Science, Honours in Statistics

at

Dalhousie University
Halifax, Nova Scotia
April 2024

Table of Contents

Chapter 1

Introduction

Determining the relationship between variables is a common goal when analyzing real data. Often a collection of covariates is used in an attempt to explain the variance in some response variable. It may be unclear which covariates have a significant effect, so estimation methods can be used to help with this. A common type of estimation, known as ordinary least squares (OLS) estimation, may be unsuitable depending on the exact setting, so we explore a different estimator in this thesis.

The alternative estimator explored in this paper is based off of Kendall's τ rank correlation. Kendall's τ is a rank based measure of association, so it only uses the ordering of the data instead of the values. Hopefully, this less specific estimator is better at detecting significance within a broad category of possible relationships. The τ based estimator will be compared to the OLS estimator in order to empirically measure which performs better and under which settings.

Chapter two of this thesis explains the necessary background information to understand the estimator used here. Chapter three explores the necessary statistical links for the τ-based estimator to function. This chapter also provides the simulation study used for measuring performance. Chapter four displays the results of testing both the τ-based estimator and the OLS estimator. Finally, chapter five includes a brief discussion of the results found in chapter four.

Chapter 2

Background

In this chapter the concepts necessary to understand the τ-based estimator are explained. This includes linear models, OLS estimation, Kendall's rank correlation, the probability integral transform, copula regression models, and bootstrapping.

2.1 Linear Models and Ordinary Least Squares Estimation

Foundational to the methods used in this paper are linear models and ordinary least squares regression. Linear models are a common form of covariate-response relationship. They take the form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$. These models are linear because the parameters of interest $\beta_0, \ldots, \beta_p$ are linearly applied to the transformed data [12, pg 212].

One can estimate the β_j for $j = 1, \ldots, p$ using ordinary least squares (OLS) estimation. Ordinary least squares estimation seeks to find the $\boldsymbol{\beta}^T = \begin{bmatrix} \beta_0 & \beta_1 & \ldots & \beta_p \end{bmatrix}^T$ parameters which minimize the total sum of squared error between the fit and the data. In other words, it seeks to minimize this quantity

$$\boldsymbol{\epsilon_i^2} = \sum_{i=1}^{n} \{y_i - (\beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})^2\}$$

Expressing the data in matrix form we have

$$Y = \boldsymbol{\beta X} + \boldsymbol{\epsilon}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

It can be shown that the $\boldsymbol{\beta}$ which minimizes the sum of squared error is $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Y$ [5, pg 18]. Where, $\boldsymbol{X}^T$ is the transpose of $\boldsymbol{X}$ and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is the matrix such that

$(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X}) = \boldsymbol{I}$ where $\boldsymbol{I}$ is the $p \times p$ identity matrix. The succinct solution to minimizing square error make it an ideal loss function for many regression scenarios, including those in this paper.

2.2 Kendall's Rank Correlation Coefficient

Central to the regression technique investigated in this thesis is the Kendall's τ rank correlation coefficient, known also as the Kendall's τ coefficient, or simply the τ correlation. It is a measure of ordinal association between two random variables X and Y with paired observations $(x_i, y_i),\ \ i = 1, \ldots, n$. One can interpret the τ correlation as the probability of concordant pairs minus the probability of discordant pairs. Concordance occurs when observations change in sync, that is, $x_i > x_j$ and $y_i > y_j$ or vice versa. Otherwise the pair is said to be discordant [8].

Suppose we have a random sample of paired measurements $(x_1, y_1)\ , \ldots, (x_n, y_n)$. To calculate the τ correlation we require the number of concordant pairs C and discordant pairs D. Kendall's τ is found via the following

$$\tau_{xy} = \frac{C - D}{\binom{n}{2}}$$

where $\binom{n}{2}$ is the total number of combinations. Recognizing that this formula is analogous to the classical formula for probability $P(A) = \frac{\text{Number of Events such that A occurs}}{\text{Total number of Events}}$ reveals the earlier interpretation of the τ coefficient. [12, pg 355,356].

In the context of this thesis, we have the simulated data $(\tilde{Y}, \tilde{\boldsymbol{X}})$ where the Kendall's τ rank correlation coefficient is calculated for each pair of variables. This results in a $(p+1) \times (p+1)$ matrix

$$\boldsymbol{K} = \begin{bmatrix} 1 & \tau_{X_1,X_2} & \tau_{X_1,X_3} & \cdots & \tau_{X_1,X_p} & \tau_{X_1,Y} \\ \tau_{X_2,X_1} & 1 & \tau_{X_2,X_3} & \cdots & \tau_{X_2,X_p} & \tau_{X_2,Y} \\ \tau_{X_3,X_1} & \tau_{X_3,X_2} & 1 & \cdots & \tau_{X_2,X_p} & \tau_{X_3,Y} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \tau_{X_p,X_1} & \tau_{X_p,X_2} & \tau_{X_p,X_3} & \cdots & 1 & \tau_{X_p,Y} \\ \tau_{Y,X_1} & \tau_{Y,X_2} & \tau_{Y,X_3} & \cdots & \tau_{Y,X_p} & 1 \end{bmatrix}$$

Where $\tau_{X_j,X_k} = \tau_{X_k,X_j}$ is the τ correlation coefficient between variables $\tilde{X}_j$ and $\tilde{X}_k$ which may include the response Y. Note how the diagonal elements are all 1 since

the τ correlation between a variable and itself is always 1. This matrix can further be split into sub matrices $\boldsymbol{K_{XX}}$ and $\boldsymbol{K_{XY}}$ viz.

$$\boldsymbol{K_{XX}} = \begin{bmatrix} 1 & \tau_{X_1,X_2} & \tau_{X_1,X_3} & \dots & \tau_{X_1,X_p} \\ \tau_{X_2,X_1} & 1 & \tau_{X_2,X_3} & \dots & \tau_{X_2,X_p} \\ \tau_{X_3,X_1} & \tau_{X_3,X_2} & 1 & \dots & \tau_{X_2,X_p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tau_{X_p,X_1} & \tau_{X_p,X_2} & \tau_{X_p,X_3} & \dots & 1 \end{bmatrix} \qquad \boldsymbol{K_{XY}} = \begin{bmatrix} \tau_{X_1,Y} \\ \tau_{X_2,Y} \\ \tau_{X_3,Y} \\ \vdots \\ \tau_{X_p,Y} \end{bmatrix}$$

The function corKendall from the `R` package `copula` was used for access to the much faster τ calculation [10].

2.3 Probability Integral Transform

An important result used to transform variables for use in copulas is the probability integral transform. Given a continuous random variable X with cumulative distribution function F_x the transformed variable $Y = F_x[X]$ has the standard uniform distribution. [2, pg 54-55]. In addition, the inverse probability integral transform of a distribution can be used to generate said distribution from the standard uniform, so for $U \sim \text{Uniform}\boldsymbol{(0,1)}$ we have $F_x^{-1}(U) \sim F_x$. It is also relevant to note that the probability integral transform for continuous distributions is strictly monotonic [3, pg 28].

2.4 Copulas

The modelling used here involves joint density functions between variables which may not be independent. The copula is used to describe these dependence structures conveniently. Technically, the copula is a type of joint cumulative distribution whose marginal distributions are standard uniform distributions. Standard uniform distributions are special in that any other distribution can transformed to one through the probability integral transform. This property is formalized in Sklar's theorem, which also states that combining a copula and arbitrary marginal distributions always creates a valid joint distribution [14, pg 229-231].

There are other methods to describe dependence structure between random variables but all of these suffer from their own issues. One could simply assume the

joint distribution between variables, but joint distributions uniquely determine the marginal distributions. Given a joint density $P(X_1, X_2)$ we find the marginal density of X_1 by integrating out the other variable(s)

$$\int_a^b P(X_1 = x_1, X_2 = x_2)dx_2$$

where a and b are the lower and upper bounds of density along the X_2 dimension. This is a definite integral, and as such has a unique solution [6, pg 314]. If the marginal distributions are to be varied, the joint distribution may need to change as well. Another technique is to use conditional distributions. In the two variable case, the conditional distribution can be combined with the marginal distribution as per this formula $P(X = x, Y = y) = P(X = x|Y = y)P(Y = y)$ [12, pg 375]. This describes the resulting joint density function, and thus the dependence structure, but the conditional $P(X = x|Y = y)$ will be dependent on the marginal distribution of Y. This allows for the marginal distribution of Y to be chosen freely, but not that of X. Compare both of these to the copula, which allows complete freedom in changing the marginal distributions while still retaining the desired dependence structure.

2.5 Bootstrapping and Confidence Intervals

In this thesis, we will use an estimator that does not have an analytically derived confidence interval. Luckily, we can use the non-parametric bootstrap to approximate the sampling distribution of such an estimator. Bootstrapping is a resampling technique used to artificially create pseudo samples from the population distribution.

Suppose $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ is a random sample of size n from some unknown distribution F. A bootstrap sample is obtained by taking a random sample of size n with replacement from the original data $\boldsymbol{x}$. For each of B bootstrap samples collected, where B is often more than 2000, the statistic of interest, say $\hat{\tau}$, is calculated. The collection of the statistic values are known as bootstrap replicates. Assigning a weight of $\frac{1}{B}$ to each bootstrap replicate yields an approximation of the sampling distribution of the statistic $\hat{\tau}$. As the sample size n increases, it can be shown that the distribution of $\hat{\tau}_1, \ldots, \hat{\tau}_B$ calculated from their corresponding bootstrap samples, will approach the true sampling distribution of $\hat{\tau}$ under certain conditions [4, pg 159]. By finding the $(\alpha/2)$ and $(1 - \alpha/2)$ quantiles, a $(1 - \alpha)\%$ confidence interval is constructed for τ

Chapter 3

Methodology

Here, we explain the τ-based estimation method of particular interest in this thesis as well as the simulation study used to test it.

3.1 Gaussian copula regression model

Copulas may be used to define models which assume a particular correlation structure but have arbitrary marginal distributions. We use one such model here, known as the Gaussian copula model. Suppose the we have observed data $(Y, \boldsymbol{X}) = (Y, X_1, X_2, \ldots, X_p)$. These variables are transformations of the true model covariates and response $(\tilde{Y}, \tilde{\boldsymbol{X}})$ which have been passed through unknown monotonic transformations $f_1(\tilde{Y}), f_2(\tilde{X}_1), f_3(\tilde{X}_2), \ldots, f_{p+1}(\tilde{X}_p)$. The unobserved $(\tilde{Y}, \tilde{\boldsymbol{X}})$ have the linear relationship

$$\tilde{Y}_i = \tilde{\boldsymbol{X}}_i^T \tilde{\boldsymbol{\beta}} + \epsilon_i, \; i = 1, 2, \ldots, n.$$

They also follow a MVN$(\mathbf{0}, \boldsymbol{\Sigma})$ where Σ is a $(p+1) \times (p+1)$ covariance matrix such that diagonal elements are 1 for identifiability [1]. The error term ϵ_i follows an unrelated Normal$(\mathbf{0}, \boldsymbol{\sigma_\epsilon})$ distribution.

We have the goal of estimating the $\tilde{\boldsymbol{\beta}}$ for variable selection. Usually, this can be accomplished using the ordinary least squares estimator $\hat{\tilde{\boldsymbol{\beta}}} = (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \tilde{Y}$. This is not possible here, since we do not have access to $(\tilde{Y}, \tilde{\boldsymbol{X}})$. However, recall that the marginal transformations $f_1, \ldots, f_{p+1}$ are monotonic. This means that rank based measures of association, such as Kendall's τ, are identical on both $(\tilde{Y}, \tilde{\boldsymbol{X}})$ and $(Y, \boldsymbol{X})$. For this reason, a τ based estimator of $\tilde{\boldsymbol{\beta}}$ calculated on the observed data $(Y, \boldsymbol{X})$ also applies to the unobserved data $(\tilde{Y}, \tilde{\boldsymbol{X}})$. We will denote such an estimator as $\hat{\boldsymbol{\beta}}_\tau$.

3.2 τ based estimation of $\tilde{\beta}$

Consider the ordinary least squares estimator $\hat{\tilde{\boldsymbol{\beta}}} = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}$. This estimator is comprised of two factors $(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}$. Direct access to these values is impossible in our setting, but there is an alternative approach.

Note the sample covariance estimate

$$\hat{\sigma}_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - E[X_j])(x_{ik} - E[X_k])$$

[13]. The mean of $(\tilde{Y}, \tilde{\boldsymbol{X}})$ is known to be 0 since they follow a MVN$(\mathbf{0}, \boldsymbol{\sigma})$ distribution. This shortens the covariance estimator to

$$\hat{\sigma}_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij})(x_{ik}).$$

This sum can be expressed as the inner product of random vectors X_j and X_k

$$\tilde{X}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix} \qquad \tilde{X}_k = \begin{bmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{bmatrix}$$

to obtain

$$\hat{\sigma}_{jk} = \frac{1}{n}\tilde{X}_j^T\tilde{X}_k$$

By combining covariates into the matrix $\tilde{\boldsymbol{X}} = \begin{bmatrix} \tilde{X}_1 & \tilde{X}_2 & \dots & \tilde{X}_p \end{bmatrix}$ We can obtain covariance estimates

$$\hat{\sigma}_{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}} = \frac{1}{n}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$$

and

$$\hat{\sigma}_{\tilde{\boldsymbol{X}}\tilde{Y}} = \frac{1}{n}\tilde{\boldsymbol{X}}^T\tilde{Y}.$$

These are very similar to the two factors $(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}$ which comprise the OLS estimator of $\tilde{\boldsymbol{\beta}}$. In other words, the OLS estimator can be viewed as a function of the unscaled covariance matrices. By estimating the covariance of $(\tilde{Y}, \tilde{\boldsymbol{X}})$ using Kendall's τ, we can create the OLS estimator for the underlying dataset.

To estimate the covariance using the τ values we use the following [1, 9]:

$$\sigma_{jk} = sin\left(\frac{\pi}{2}\tau_{X_j,X_k}\right)$$

There is now a link between the tau matrix $\boldsymbol{K}$ and $\hat{\boldsymbol{\beta}}_\tau$ through the following:

$$sin\left(\frac{\pi}{2}\boldsymbol{K}_{\tilde{X}\tilde{X}}\right) = \hat{\Sigma}_{\tilde{X}\tilde{X}} = \frac{1}{n}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$$

$$sin\left(\frac{\pi}{2}\boldsymbol{K}_{\tilde{X}\tilde{Y}}\right) = \hat{\Sigma}_{\tilde{X}\tilde{Y}} = \frac{1}{n}\tilde{\boldsymbol{X}}^T\tilde{Y}$$

Where $\hat{\Sigma}_{\tilde{X}\tilde{X}}$ is the $p \times p$ matrix comprised of the covariances between covariates and $\hat{\Sigma}_{\tilde{X}\tilde{Y}}$ is the $p \times 1$ matrix comprised of the covariances between response and covariates. Multiplying the estimated covariance matrices $\hat{\Sigma}_{\tilde{X}\tilde{X}}$ and $\hat{\Sigma}_{\tilde{X}\tilde{Y}}$ by n will yield $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{X}}^T\tilde{Y}$ respectively. The ordinary least squares estimators of $\hat{\tilde{\boldsymbol{\beta}}}$ can then be directly calculated using $(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{Y}$ for our final estimates.

3.3 Variable Selection Using Bootstrapping

To check for covariate significance, we make use of the equivalence between 2-sided hypothesis tests and confidence intervals. Testing $H_0 : \tilde{\beta}_j = 0$ versus $H_a : \tilde{\beta}_j \neq 0$ is done by checking if 0 is within the confidence interval, and if it is not, we reject H_0. A rejection of the null hypothesis would imply that $\tilde{\beta}_j$ is not 0 and thus has an effect on the response [11, pg 242]. Given the sampling distribution of $\hat{\boldsymbol{\beta}}_\tau$ is unknown, a nonparametric bootstrap will be used with 3000 replicates to calculate the appropriate $100(1-\alpha)\%$ confidence intervals for $\tilde{\boldsymbol{\beta}}$.

Chapter 4

Simulations

4.1 Generating Simulated Data

To compare the effectiveness of the two estimation methods for variable significance we use simulated data. The original $\tilde{\boldsymbol{\beta}}$ parameters of the simulated data are known, so estimated $\hat{\boldsymbol{\beta}}_\tau$ can be directly compared to determine reliability. Performance may change depending on the generating settings, so multiple settings are tested to best understand when the estimators perform well. Implementation in computer code was performed using the R coding language and the Rstudio integrated development environment [7].

We first simulate $\tilde{\boldsymbol{X}}$ by generating n random samples from p Normal$\mathbf{(0,1)}$ distributions. The response variable is created by

$$\tilde{Y}_i = \tilde{\boldsymbol{X}}_i^T \tilde{\boldsymbol{\beta}} + \epsilon_i,\ i = 1, 2, \ldots, n.$$

This ensures that the underlying model fits the Gaussian copula model while providing us with the true $\tilde{\boldsymbol{\beta}}$ [11, pg 475]. To create the observed data $(Y, \boldsymbol{X})$ we use the monotonic transformations $f_1(\tilde{Y}), \ldots, f_{p+1}(\tilde{X}_j)$. Not all $\tilde{Y}, \tilde{\boldsymbol{X}}$ need to be transformed, so in some cases $\tilde{X}_j = X_j$. This is equivalent to applying the transformation $f_j(X_j) = 1 \times X_j$ which is obviously monotonic.

For comparison we also fit the linear model

$$Y_i = \boldsymbol{X}_i \boldsymbol{\beta} + \epsilon_i \quad i = 1, \ldots, n$$

and compute the standard ordinary least squares estimate $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Y$ This model is ignorant to the true relationship involving the $\tilde{Y}, \tilde{\boldsymbol{X}}$ but it is simple. If it is found that the $\tilde{\boldsymbol{\beta}}_\tau$ performs similarly to $\tilde{\boldsymbol{\beta}}_{OLS}$ for the purposes of variable selection, it may not be worth the additional effort to use the former.

One setting that varied between simulations is the transformations applied to both covariates and response. These transformations include strictly monotonic functions

such as exponential, logarithmic, hyperbolic, and polynomial as well as the probability integral transform. In some cases, the range on which a function is monotonic included only positive numbers, in which case the data was transformed using PIT to a density with only positive values beforehand. Additionally, the sample size n, the error variance, and the proportion of significant covariates were also varied. The true $\tilde{\boldsymbol{\beta}}$ were fixed at the beginning of each trial.

For each scenario there were 10 covariates with 3 being significant ($\tilde{\beta}_i \neq 0$). Additionally, each scenario was tested using 25 different random samples. Performance of both estimators was measured using sensitivity and specificity. For each model fit we can compare the detected significance of $\hat{\boldsymbol{\beta}}_\tau$ and $\hat{\boldsymbol{\beta}}_{OLS}$ against the true significance of $\tilde{\boldsymbol{\beta}}$. If a component coincides it is known as a true positive (TP) or true negative (TN). If an insignificant $\tilde{\beta}_j$ is detected as significant it is known as a false positive (FP) and if a significant $\tilde{\beta}_j$ is detected as insignificant it is known as a false negative (FN). Sensitivity and specificity are functions of the total amounts of TP, TN, FP, and FN via

$$\text{Sensitivity} = \frac{TP}{TP + FN} \qquad \text{Specificity} = \frac{TN}{TN + FP}.$$

Sensitivity can be interpreted as the ratio of significant $\tilde{\beta}_j s$ detected while specificity can be interpreted as the ratio of insignificant $\tilde{\beta}_j s$ properly detected as such. These metrics are tabulated across each scenario for both estimators.

The theoretically perfect table is the following:

Table 4.1: Ideal Table

	Template		Ideal Score	
τ	Sensitivity	Specificity	1	1
OLS	Sensitivity	Specificity	1	1

The perfect score is not obtained by any setting, instead a benchmark scenario is chosen. The benchmark setting produces the following model:

$$\tilde{Y} = \tilde{\beta}_1\tilde{X}_1 + \tilde{\beta}_2\tilde{X}_2 + \tilde{\beta}_3\tilde{X}_3 + \cdots + \tilde{\beta}_{10}\tilde{X}_{10} + \epsilon \quad \text{where} \quad \epsilon \sim N(\mathbf{0}, \mathbf{0.1})$$

$$\tilde{X}_i^T = \begin{bmatrix} x_1 & x_2 & \dots & x_{1000} \end{bmatrix}^T \quad \text{with} \quad x_i \overset{iid}{\sim} N(\mathbf{0}, \mathbf{1})$$

$$\tilde{\beta}_1 = 0.1, \quad \tilde{\beta}_2 = 0.3, \quad \tilde{\beta}_3 = 1, \quad \tilde{\beta}_{4,\dots,10} = 0$$

No marginal transformations are applied to covariates or response. This produces the following table:

Table 4.2: Standard Table without transformations

	Template		Ideal Score	
τ	Sensitivity	Specificity	1	0.96
OLS	Sensitivity	Specificity	1	0.926

Note that sensitivity is labelled in blue and specificity in green, as the template column will no longer present in tables from here on.

4.2 Distribution of Covariates $\boldsymbol{X}_i$

Table 4.3: PIT of covariates

	Normal(0,1)		$\chi^2(3)$		$\beta(5,1)$		$\Gamma(2,2)$		Exp(1.5)		MVN($\rho = 0.9$)	
τ	1	0.96	1	0.96	1	0.96	1	0.96	1	0.96	1	0.954
OLS	1	0.926	1	0.96	1	0.949	1	0.96	1	0.971	1	0.949

Performance of the OLS estimator was rather stable here with only minor dips occasionally. The τ-based estimator was invariant to the transformations since they are strictly monotonic. The multivariate normal (MVN) distribution had a covariance matrix such that all non-diagonal elements were equal to $\rho = 0.9$. The τ-based estimator performance changes under the MVN distribution due to it generating new samples from scratch as opposed to transforming existing samples.

4.3 Observation number n

Table 4.4: Change in observation number

	n = 1000		n = 100		n = 30	
τ	1	0.96	0.973	0.994	0	1
OLS	1	0.926	1	0.966	1	0.971

At $n = 1000$ and $n = 100$ both estimators performed well. The τ-based estimator had a decrease in sensitivity at $n = 100$ but it was made up for by an increase in specificity. At $n = 30$ the τ-based estimator had 0 sensitivity. This makes it unusable,

as it will never detect any significant covariates. The OLS estimator performed quite well for comparison, so it may be preferred at very low observation counts.

4.4 Covariate Transformations $\boldsymbol{X}_i$

Table 4.5: Marginal Transformations

	e^x		$log_e(x)$		$1/(-x)$		x^2	
τ	1	0.96	1	0.96	1	0.96	1	0.96
OLS	0.507	0.903	1	0.949	0.853	0.971	0.973	0.96

Performance of the OLS estimator was satisfactory under the logarithmic and quadratic transformations but decreased moderately at the hyperbolic transformation. Most notably, the exponential transformation greatly decreased the OLS estimator's sensitivity to 0.507. The τ-based estimator was invariant under these transformations so it performed quite well still. Covariates were first transformed to a χ^2_3 distribution using PIT so that the range of values is non-negative since x^2 and $1/(-x)$ are monotonic on $x_i \in (0, \infty)$.

4.5 Response Transformations $\boldsymbol{Y}$

Table 4.6: Transformation applied to Response

	e^y		$log_e(y)$		$1/(-y)$		y^2	
τ	1	0.96	1	0.96	1	0.96	1	0.96
OLS	0.4	0.937	1	0.943	0.827	0.966	0.973	0.949

This section has similar results to that of the covariate transformations. The OLS model performs moderately poorly under a hyperbolic response transformation and extremely poorly under an exponential response transformation. The τ-based model was invariant once again, producing solid results. Response data was first transformed to χ^2_3 to restrict the range to only positive numbers.

4.6 Error Variance

Table 4.7: Changes in σ_ϵ

	0.1		5		20	
τ	1	0.96	0.507	0.926	0.187	0.926
OLS	1	0.926	0.493	0.926	0.2	0.926

The results here are uninteresting, as the error variance increases both estimators cease to function well. The error is applied to the underlying model directly, so an increase in variance will decrease the strength of the underlying fit. The estimators here are simply detecting that decrease.

4.7 Proportion of Significant Covariates

For this section 30 covariates are used. In the "$> 10\%$" column, 9 covariates are significant out of 30 and in the "$= 10\%$" column, 3 covariates are significant out of 30.

Table 4.8: Proportion of $\tilde{\beta}_j \neq 0$

	$> 10\%$		$= 10\%$	
τ	1	0.964	1	0.967
OLS	1	0.962	1	0.954

Performance of both estimators varies little here. The proportion of covariates which are significant seems to have little effect on performance.

4.8 Conditional Distribution $Y|\boldsymbol{X}$

Table 4.9: PIT applied to Response

	Normal(0,1)		$\chi^2(3)$		$\beta(5,1)$		$\Gamma(2,2)$		Exp(1.5)	
τ	1	0.96	1	0.96	1	0.96	1	0.96	1	0.96
OLS	1	0.926	1	0.949	1	0.943	1	0.954	1	0.954

For this section the PIT was applied to only the response variable without any additional monotonic transformations afterward. Performance overall is solid for both

estimators, much like the when the PIT was applied to the covariates. The OLS estimator varies slightly depending on the exact distribution of $(Y|\boldsymbol{X})$ but sensitivity remains at 1 and specificity never drops below 0.92, so it is certainly still usable.

Chapter 5

Discussion

There is a notable conceptual advantage the τ-based estimator has for variable selection which motivated testing. Kendall's τ is a rank based measure of association, as such different datasets may produce identical τ values so long as the ordering of their observations remains the same. Many common variable transformations, such exponential or polynomial are among those that do not change the ordering of observations. The τ-based estimator can be used to test the significance of all such transformations simultaneously, allowing for a large class of relationships to be marked for further analysis.

Despite this supposed advantage the τ-based estimator has, we can use the results from the simulation study to empirically compare its performance against the OLS estimator. Varying the distribution of covariates, conditional distribution of the response, proportion of significant covariates, and error variance had uninteresting results. The first three scenarios produced similar, steady performance in both the τ-based estimator and the OLS estimator. Increasing the error variance decreased the performance of both estimators to an even degree. Choosing between these two estimators is arbitrary under these 4 scenarios.

The first of the significant differences occured when we varied the monotonic transformations applied to the covariates. The specific transformations we tested were e^x, $log_e(x)$, $1/(-x)$ and x^2. For each of these transformations the PIT was used first to guarantee that the resulting data would be non-negative. For the logarithmic and quadratic transformations both estimators performed similarly but for the hyperbolic transformation the OLS estimator noticeably dipped in sensitivity. Ordinarily, the sensitivity is for the OLS estimator is 1 but under the hyperbolic transformation $1/(-x)$ it becomes 0.853. This is second to the massive decrease in performance the OLS estimator undergoes during the exponential transformation e^x. Under this transformation, the OLS estimator's sensitivity drops to just 0.507. It is detecting

false negatives nearly more often than true positives here. Additionally, the specificity in this scenario is 0.903, one of the worst we have seen so far from the OLS estimator.

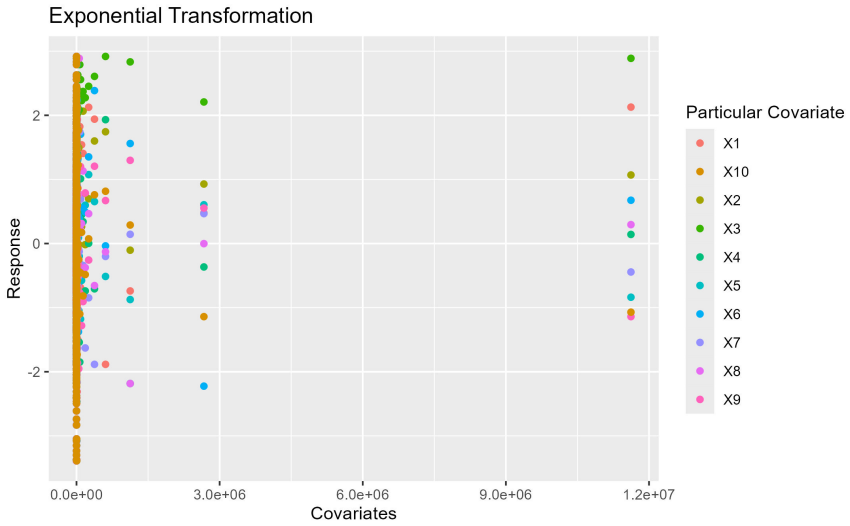

Figure 5.1: Scatterplot of Exponential Covariates

Figure 5.2: Scatterplot of Hyperbolic Covariates

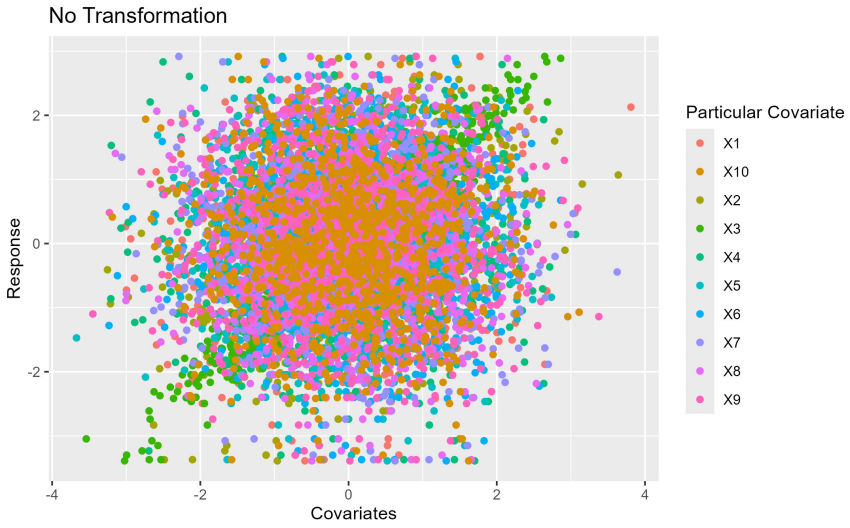

Figure 5.3: No Marginal Transformations

From the figures above, it would seem that the squishing of most covariate observations toward a particular value is the cause of the poor performance under the hyperbolic and exponential transformations. This is in contrast to the data which did not undergo any transformations that is more spread out where the OLS estimator performed much better.

Clearly, if there is a possibility that an exponential relationship links covariate and response this could greatly reduce confidence in the OLS estimator's abilities. This also extends to transformations applied to the response, where the OLS estimator performs similarly poorly when the transformation e^y has been applied. In either case the τ-based estimator is invariant in its performance, obtaining sensitivity of 1 and specificity of 0.96.

There is another scenario where performance differed greatly, this time against the τ-based estimator. While varying the sample size n it was found that at $n = 1000$ and $n = 100$ the τ-based estimator and OLS estimator had approximately the same sensitivity and specificity. However, at $n = 30$ the τ-based estimator had an abysmal sensitivity of 0. With a specificity of 1 it is clear the estimator indicated every single covariate to be insignificant. For comparison, the OLS estimator had a sensitivity of 1 and a specificity of 0.971. To understand why the τ estimator performed so poorly we can look at a histogram of the estimated $\hat{\tilde{\beta}}_3$ replicates. The third covariate was chosen because it had the greatest true $\tilde{\beta}_j$ value at $\tilde{\beta}_3 = 1$.

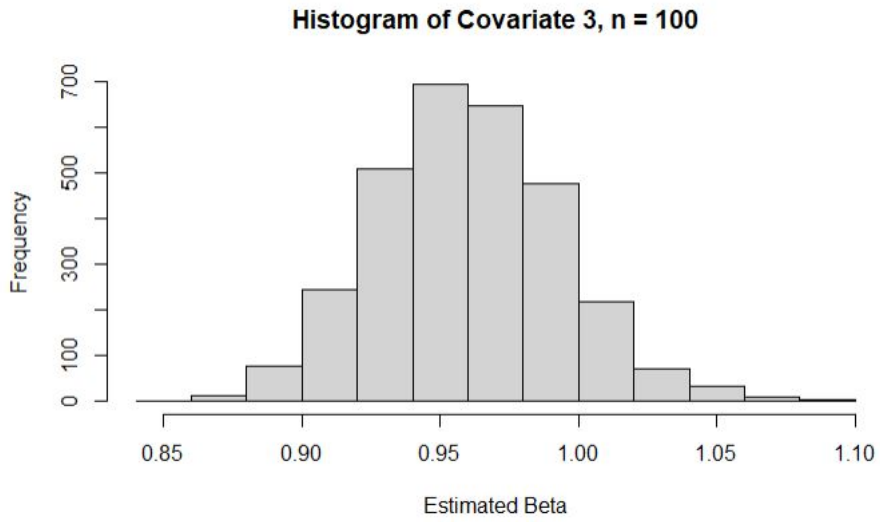

Figure 5.4: Histogram of Bootstrap Replicates at $n = 100$

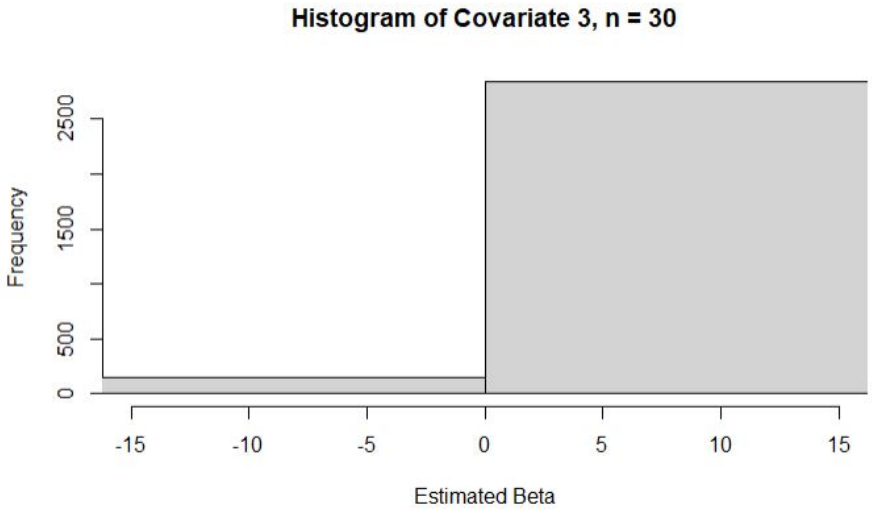

Figure 5.5: Histogram of Bootstrap Replicates at $n = 30$

These histograms reveal that the poor performance of the τ-based estimator is caused by the bootstrap. At low observation counts, the bootstrap has a tendency to classify $\hat{\tilde{\beta}}_j$ replicates as insignificant despite not being so. If an exact formula for the standard error of the τ-based estimator was found perhaps this could be solved.

Taking the results together, we can make criteria for deciding when the τ-based estimator should be preferred and when the OLS estimator should be preferred. If the

number of observations is low, at less than 100, the OLS estimator is more effective. Due to the bootstrap, the τ-based estimator greatly struggles to detect significant covariates at $n \leq 30$. If the number of observations is adequate, we then determine whether the response variable may have an exponential relationship with the covariates. If there is a possibility this is the case, the τ estimator should be preferred. The OLS estimator had a large number of false negatives and false positives when response was an exponential function of covariates. If the data has both high observation count and there is no possibility of response being an exponential function of covariates, then both estimators perform equally well.

5.1 Conclusion

Despite the supposed conceptual strengths of the τ-based estimator for determining which covariates have a significant effect on the response, empirical evidence illustrated how well the naive OLS estimator kept up. The only scenario it should be avoided was when an exponential transformation linked response and covariates. It could be found that additional transformations cause the OLS estimator to fail, but based on the results here it may still be preferable to continue using the OLS estimator for variable selection. The OLS estimator is well understood and relatively simple to implement, compared to the relatively recent development of copula models and this τ based estimator. Choosing the τ-based estimator forces it to compete with the decades of use of the OLS estimator. This familiarity may be preferable to the minor gains in performance of a relatively undeveloped estimator.

Bibliography

[1] T Tony Cai and Linjun Zhang. High-dimensional gaussian copula regression: Adaptive estimation and statistical inference. *Statistica Sinica*, pages 963–993, 2018.

[2] George Casella and Roger L. Berger. *Statistical Inference Second Edition*. Duxbury, 2002.

[3] Luc Devroye. Nonuniform random variate generation. *Handbooks in operations research and management science*, 13:83–121, 2006.

[4] Bradley Efron and Trevor Hastie. *Computer age statistical inference, student edition: algorithms, evidence, and data science*, volume 6. Cambridge University Press, 2021.

[5] Fumio Hayashi. *Econometrics*. Princeton University Press, 2000.

[6] Saleem Watson James Stewart, Daniel Clegg. *Single Variable Calculus*. Cengage, 2021.

[7] colleagues John Chambers. R. 2024.

[8] Maurice George Kendall and Jean Dickinson Gibbons. *Rank correlation methods.* Oxford University Press, 1990.

[9] William H. Kruskal. Ordinal measures of association. *Journal of the American Statistical Association*, 53(284):814–861, 1958.

[10] Johanna G. Nešlehová Rebecca Morger Marius Hofert, Ivan Kojadinovic. copula: Multivariate dependence with copulas. 2023.

[11] Kjell Doksum Peter Bickel. *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice Hall, 2001.

[12] David Bock Augustin Vukov Augustine Wong Richard DeVeaux, Paul Velleman. *Stats: Data and Models*. Pearson, 3rd edition, 2019.

[13] Dean Wichern Richard Johnson. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007.

[14] Abe Sklar. *Fonctions de répartition à n dimensions et leurs marges*. Paris Institute of Statistics, 1959.