# 3. A COMMENT ON "A REVIEW OF STUDENT TEST PROPERTIES IN CONDITION OF MULTIFACTORIAL LINEAR REGRESSION"

Eric EISENSTAT*

## Abstract

*A recent article (Pavelescu, 2009) proposes a correction to the conventional student-t test of significance in linear regression models, but offers no formal description of its properties. This comment formally characterizes the sampling properties of the corrected student-t statistic. In application to multifactorial regressions, it turns out that the corrected student-t statistic is not ancillary – its sampling distribution depends on unknown nuisance parameters.Therefore, it is impossible to reasonably compute critical values and operatively designate a rejection criterion using such a test statistic, which makes the proposed testing procedure impractical. Some suggestions regarding the search for similar testing procedures are proposed and a Bayesian alternative is further discussed.*

**Keywords:** multifactorial classical normal regression, collinearity, multicollinearity, significance test, sampling distributions, power functions, Bayesian linear regression, prior information, posterior distributions

**JEL Classification**: C11, C12

### 1. Multifactorial Regression and Collinearity

Collinearity and its exotic counterpart *multicollinearity* are well explored topics in econometric literature (for an extensive overview and analysis, see [Judge, Hill, Griffiths, Lütkepohl and Lee, 1988, pp. 859-881)]. The fundamental problem for econometricians is that strong linear relationships amongst explanatory variables pose not only a formidable impedance on statistical inference regarding individual parameters, but a vastly elusive one at that, since economists rarely have direct control over the *data generating process*. From a statistics perspective, the common view projects that collinearity undermines accurate inference through its effect on the standard errors of individual parameter estimates: *ceteris paribus*, stronger collinearity proportionately increases standard errors, leads to wider confidence intervals, and lower test statistics (in absolute value) in significance tests.[1]

* *Department of Economics, University of California, Irvine, United States, e-mail: eric.eisenstat@gmail.com.*

Countless methods of “detecting” multicollinearity and containing its effects *ex post* have been proposed. In lieu of redesigning experiments that generate the data or obtaining larger samples (options which are most often simply not available to econometricians), the operational “solutions” in the literature almost exclusively focus on either the systematic inclusion/exclusion of certain explanatory variables, or the reconditioning of explanatory variables such as to induce orthogonality and yield lower degrees of collinearity according to some predetermined measure, depending on a particular case of interest. From this viewpoint, the proposal of a different test statistic in (Pavelescu, 2009) to *generally* mitigate multicollinearity in linear regression significance testing offers a potentially interesting new perspective on an exhausted subject. More specifically, (Pavelescu, 2009) formulates a *corrected* student-t statistic which is claimed to aid the researcher in countering the effects of collinearity on assessing the “relevance” of estimated parameters. This claim, however, is scarcely justified in the traditional sense of hypothesis testing analyses as it lacks an in depth discussion of any properties beyond its algebraic relationship to the *standard student-t* and *F* statistics.

It must be emphasized that in order to understand how such a proposed testing procedure is optimally applied, a diligent researcher requires a rigorous examination of its theoretical properties, especially in comparison with already existing and commonly accepted techniques. This often involves a comparison of sampling distributions, critical regions, power functions, etc. Of course, the primary property that *any* viable test statistic must exhibit is that it is *ancillary* under the null hypothesis, meaning that its null sampling distribution does not depend on unknown parameters. The need for this requirement is clear: a test statistic that is not ancillary in its sampling properties fails to achieve any sense of *standardization*, which is always the foremost goal of its application, and consequently, its magnitude is of no practical meaning.

Unfortunately, the *corrected* student-t statistic is exactly of such non-ancillary nature under the null hypothesis for any regression with two or more explanatory variables, and this fact is attested to formally in the following section. Moreover, it is not directly obvious how this statistic might be restructured to generate an ancillary statistic (even one that is not immediately operational), and therefore, it is impossible to utilize the *corrected* student-t statistic in constructing critical regions, as applicable to multifactorial regressions. It is further shown that in the case of a unifactorial regression (with exactly one explanatory variable), employing the *corrected* static to derive a critical region in the manner implied by the text results in a testing procedure that is equivalent to the traditional student-t test.

To facilitate the ensuing exposition, the remainder of this section is dedicated to a basic overview of the distribution theory relevant to the multifactorial *Classical Normal Regression (CNR)* model with emphasis on the properties that are of particular interest to our discussion. Hence, recall that the multifactorial regression postulates a

*[1] This is, however, not generally true for linear combinations of parameters. For an example where multicollinearity leads to increased power in a hypothesis test on a sum of regression parameters, see (Goldberger, 1991, pp. 250-251).*

linear relationship between a dependent variable $y$ and $k$ explanatory variables $x_1, \dots, x_k$ of the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \sum_{j=2}^{k} \beta_j x_{ij} + \epsilon_i \tag{1}$$

where in accordance with the assumptions of the CNR model, $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.
Without loss of generality, we designate $\beta_1$ to be the parameter of interest and $\beta_0, \beta_2, \dots, \beta_k$ and $\sigma^2$ to be the *nuisance* parameters. Moreover, *we* will denote the *Ordinary Least Squares (OLS)* estimators of $\beta_j$, and $\sigma^2$ with $\hat{\beta}_j$, and $\hat{\sigma}^2$, respectively.

In what follows, it is convenient to express the estimators employing the following matrix notation: let $\boldsymbol{y}$ be the $n \times 1$ vector consisting of dependent variables $\boldsymbol{y} = (y_1 \cdots y_n)'$, $\hat{\boldsymbol{\beta}}$ the $(k+1) \times 1$ vector $(\hat{\beta}_0 \cdots \hat{\beta}_k)'$, and let $\boldsymbol{X}$ denote the $n \times (k+1)$ matrix with elements $X(i,1) = 1$ and $X(i,j) = x_{i,j-1}$ for $i = 1, \dots, n$ and $j = 2, \dots, k+1$. Furthermore, define $\boldsymbol{Q} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$, with $\boldsymbol{q}_j$ representing the $j^{th}$ row of $\boldsymbol{Q}$ and $q_{jh}$ denoting the element of $\boldsymbol{Q}$ located at row $j$ and column $h$. Then,

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-k-1} \tag{2}$$

$$\hat{\beta}_1 = \boldsymbol{q}_2 \boldsymbol{X}'\boldsymbol{y} \tag{3}$$

It is well known that $\hat{\beta}_1$ and $\hat{\sigma}^2$ are stochastically independent and marginally follow the distributions

$$(n-k-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-k-1) \tag{4}$$

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma^2 q_{22}) \tag{5}$$

By letting $R^2_{1,k}$ represent the *coefficient of determination* obtained from regressing $x_1$ on all other explanatory variables $x_2, \dots, x_k$, and $s^2_{x_1}$ the sample variance of $x_1$, we may write $q_{22} = \frac{1}{n(1-R^2_{1,k})s^2_{x_1}}$. From this expression, the conventional view on the effect of multicollinearity is immediately evident: as collinearty between $x_1$ and other explanatory variables increases, $R^2_{1,k} \to 1$ and for a fixed sample size $\mathrm{var}(\hat{\beta}_1) \to \infty$. Thus, multicollinearity is typically related to imprecise estimates, which is in turn, reflected in wide confidence intervals as well as weak *power* in hypothesis tests on individual parameters.
The standard *student-t test* of significance, which is synonymous with the hypothesis test

$$H_0: \beta_1 = 0, \quad H_A: \beta_1 \neq 0$$

proceeds by constructing the *student-t statistic*

$$\hat{t}_1 = \frac{\hat{\beta}_1/\sqrt{q_{22}}}{\hat{\sigma}^2} \qquad (6)$$

and defining the rejection rule

$$\Re_1 : \ reject\ H_0 \text{ if } |\hat{t}_1| \geq t_\alpha$$

where the critical value $t_\alpha$ is by design a function of the significance level $\alpha$ and is computed in a way that insures the *probability of a Type I error* (e.g. probability of proclaiming a parameter “statistically significant” when it is in fact zero) occurring is no greater than $\alpha$. The particular formula to determine $t_\alpha$ for a desired significance level is predicated on the fact that by (4) and (5), the student-t statistic $\hat{t}_1$ under the null hypothesis follows the *student-t distribution* with $n-k-1$ degrees of freedom:

$$\hat{t}_1 \sim t(n-k-1) \qquad (7)$$

and hence the probability of Type I error is

$$\Pr(|\hat{t}_1| \geq t_\alpha \,|\, H_0) = 2(1-\Psi_{n-k-1}(t_\alpha)) \qquad (8)$$

where $\Psi_{n-k-1}(\cdot)$ denotes the *cumulative distribution function (cdf)* of the *student-t* distribution with $n-k-1$ degrees of freedom.

Thus, the critical value $t_\alpha$ is straightforwardly computed by equating $2(1-\Psi_{n-k-1}(t_\alpha)) = \alpha$, and is practically feasible due to the fact that this expression contains only the *known* quantities $\alpha$ and $n-k-1$.

## 2. Properties of the *corrected* student-t test

Now, consider the *corrected* student-t testing procedure proposed in (Pavelescu, 2009). The specific test statistic, termed *corrected* student-t statistic, is explicitly provided in equation (19) of (Pavelescu, 2009) as

$$\hat{\tau}_1 = \hat{\lambda}_1 \hat{t}_1 = \hat{\lambda}_1 \frac{\hat{\beta}_1/\sqrt{q_{22}}}{\hat{\sigma}^2} \quad \text{with} \quad \hat{\lambda}_1 = \frac{|r_{y.x_1}|}{r_{y.x_1}} \qquad (9)$$

where $r_{y.x_1}$ denotes the sample correlation between $y$ and $x_1$.

For our purposes, it will be more convenient to write $\hat{\lambda}_1$ as

$$\hat{\lambda}_1 = \begin{cases} -1 & \text{if } \hat{\alpha}_1 < 0 \\ 1 & \text{if } \hat{\alpha}_1 > 0 \end{cases} \qquad (10)$$

where $\hat{\alpha}_1 = \frac{s_{y.x_1}}{s^2_{x_1}}$ (with $s_{y.x_1}$ denoting the sample covariance between $y$ and $x_1$) is the OLS slope estimator obtained by regressing $y$ on $x_1$ only.[2]

[2] *Note that expressing $\hat{\lambda}_1$ this way does not affect the functional definition. For completeness, it would be more appropriate to define*

Aside from the test statistics, (Pavelescu, 2009) offers no formal account of the remaining steps in completing the proposed hypothesis test (rule for rejecting the null hypothesis, computation of the critical region, etc.). We deduce from the discussion in the text, that the author intends the rejection rule to be[3]

$$\Re_2 : \textit{reject } H_0 \text{ if } \hat{\tau}_1 \geq \tau_\alpha$$

Let us first consider the properties of this testing procedure in application to unifactorial linear regressions, where $\hat{\beta}_1 = \hat{\alpha}_1$. Since in this case $\hat{\alpha}_1$ and $\hat{t}_1 = \frac{\hat{\alpha}_1 s_{x_1}\sqrt{n}}{\hat{\sigma}^2}$ always have the same sign (e.g. $\hat{\alpha}_1 < 0$ if and only if $\hat{t}_1 < 0$), it is clear that $\hat{\tau}_1 = |\hat{t}_1|$. Moreover, applying $\Re_2$ on $\hat{\tau}_1$ is equivalent to applying $\Re_1$ on $\hat{t}_1$, and hence, $\tau_\alpha = t_\alpha$ for the same significance level $\alpha$. Consequently, the two tests are equivalent in unifactorial regressions.

Of course, we are more interested in the general case with $k \geq 2$ explanatory variables, where the relationship between $\hat{\tau}_1$ and $\hat{t}_1$ is now more complex and the null sampling distribution for $\hat{\tau}_1$ must be derived explicitly in order to allow for a proper examination of its properties. This distribution is obtained through a *change of variable* technique on the transformation $(\hat{\beta}_1, \hat{\sigma}^2, \hat{\lambda}_1 \to \hat{\tau}_1)$ given in (9)-(10), where the sampling distributions of $\hat{\beta}_1$ and $\hat{\sigma}^2$ are provided in (4)-(5), and the distribution of $\hat{\lambda}_1 \mid \hat{\beta}_1, \hat{\sigma}^2$ is derived as follows:

1. Note that for the $k$-factorial regression, the coefficient estimate $\hat{\beta}_1$ may be expressed as $\hat{\beta}_1 = \hat{\alpha}_1 - \sum_{j=2}^{k} \hat{\delta}_{1j}\hat{\beta}_j$ with $\hat{\delta}_{1j} = \frac{s_{x_j,x_1}}{s_{x_1}^2}$ representing the OLS slope estimator obtained by regressing $x_j$ on $x_1$. Therefore, $\hat{\alpha}_1 < 0$ if and only if $\sum_{j=2}^{k} \hat{\delta}_{1j}\hat{\beta}_j < -\hat{\beta}_1$.
2. Denoting $\hat{b}_{1,k} = \sum_{j=2}^{k} \hat{\delta}_{1j}\hat{\beta}_j$, it is straightforward to show that conditional on $\hat{\beta}_1$, the statistic $\hat{b}_{1,k}$ is independent from $\hat{\sigma}^2$ and follows the distribution

$$\hat{b}_{1,k} \mid \hat{\beta}_1 \sim \mathcal{N}\left(b_{1,k} - R_{1,k}^2(\hat{\beta}_1 - \beta_1), \frac{\sigma^2 R_{1,k}^2}{s_{x_1}^2 n}\right) \qquad (11)$$

   where $b_{1,k} = \sum_{j=2}^{k} \hat{\delta}_{1j}\beta_j$ is the *population analogue* of $\hat{b}_{1,k}$.
3. Accordingly, the probability of interest $\Pr\left(\hat{\lambda}_1 = -1 \mid \hat{\beta}_1, \hat{\sigma}^2\right)$ under the null hypothesis $\beta_1 = 0$ is derived as

$$\hat{\lambda}_1 = \begin{cases} -1 & \textit{if} \quad \hat{\alpha}_1 \leq 0 \\ 1 & \textit{if} \quad \hat{\alpha}_1 > 0 \end{cases}$$

*in order to accommodate the special case when $\hat{\alpha}_1 = 0$ (or equivalently $r_{y,x_1} = 0$). However, since this exerts no practical impact on neither the discussion in (Pavelescu, 2009) nor the present text, we prefer to ignore the minor theoretical deficiency in favor of remaining consistent with the original definition in (Pavelescu, 2009).*

[3] *For example, see the discussion in section 2 of (Pavelescu, 2009).*

$$\Pr\left(\hat{\lambda}_1 = -1 \mid \hat{\beta}_1, \hat{\sigma}^2\right) = \Pr\left(\hat{b}_{1,k} < -\hat{\beta}_1 \mid \hat{\beta}_1\right) = \Phi\left(-\frac{s_{x_1}\sqrt{n}}{\sigma R_{1,k}}\left(b_{1,k} + \left(1 - R_{1,k}^2\right)\hat{\beta}_1\right)\right) \quad (12)$$

where $\Phi(\cdot)$ denotes the *cdf* of the *standard normal* distribution. Note that the above equation fully specifies the *conditional* sampling distribution of $\hat{\lambda}_1$ since $\Pr\left(\hat{\lambda}_1 = 1 \mid \hat{\beta}_1, \hat{\sigma}^2\right) = 1 - \Pr\left(\hat{\lambda}_1 = -1 \mid \hat{\beta}_1, \hat{\sigma}^2\right)$.

4. Now, applying the change of variable yields the sampling distribution of $\hat{\tau}_1$:

$$\hat{\tau}_1 \sim \int_0^{\infty} \left[1 + \tilde{\Phi}(-\hat{\tau}_1\omega) - \tilde{\Phi}(\hat{\tau}_1\omega)\right]\phi\left(\frac{\hat{\tau}_1\omega}{\sqrt{n-k-1}}\right)p_{n-k-1}(\omega^2)\,d\omega^2 \quad (13)$$

where $\phi(\cdot)$ denotes the *probability density function (pdf)* of the standard normal distribution, $p_{n-k-1}(\cdot)$ is the *pdf* of the $\chi^2$ distribution with $n-k-1$ degrees of freedom, and $\tilde{\Phi}(z) = \Phi\left(-\frac{s_{x_1}\sqrt{n}}{\sigma R_{1,k}}b_{1,k} - z\sqrt{\frac{1-R_{1,k}^2}{R_{1,k}^2}}\right)$.

Observe that since $\tilde{\Phi}(z)$ depends on the unknown nuisance parameters $\beta_2, \dots, \beta_k$ (through $b_{1,k}$) and $\sigma^2$ the sampling distribution (13) also depends on these parameters, and more importantly, so does the probability $\Pr(\hat{\tau}_1 > \tau_\alpha \mid H_0)$. Thus, attempting to employ the *corrected* student-t statistic with rejection rule $\Re_2$ leaves us in a rather paradoxical position: if we carry out the hypothesis test by choosing an arbitrary critical value (such as $\tau_\alpha = 0$ or $\tau_\alpha = t_\alpha$, for example), then the probability of Type I error associated with such a testing procedure depends on unknown parameters, and hence, we have no way of assessing the degree to which such a test will mislead us into erroneously proclaiming a coefficient significant, when it is in fact zero. On the other hand, it is operationally impossible to compute the critical value $\tau_\alpha$ based on a desired significance level by solving the equation $1 - F_{\hat{\tau}_1}(\tau_\alpha) = \alpha$ (where $F_{\hat{\tau}_1}(\cdot)$ is the *cdf* of the distribution in (13)) since the solution will once again depend on $\beta_2, \dots, \beta_k$ and $\sigma^2$.

Clearly, either option is unreasonable in the context of statistical inference through hypothesis testing. Therefore, the *corrected* student-t statistic as formulated in (Pavelescu, 2009) is *not operational,* least of all an improvement upon the conventional student-t test in discerning "confusions between truly relevant estimated values and 'statistical illusions'" arising from collinearity in the explanatory variables.

## 3. Remarks

It is worthwhile to point out that the result of the previous section arises directly from $\hat{\lambda}_1$ being defined in terms of the coefficient estimator $\hat{\alpha}_1$, which suggests that there may exist several ways to circumvent this problem. One simple approach worth considering is the test statistic and consequent testing procedure derived by defining explicitly the distribution of $\hat{\lambda}_1$ in terms of fixed probabilities, such that $\Pr(\hat{\lambda}_1 = -1 \mid \hat{\beta}_1, \hat{\sigma}^2) = \eta$.[4] Although such a testing procedure is of little practical value, it turns out that it is theoretically interesting in the sense that its properties are in many important ways analogous to the properties of a wide variety of conceivable tests within the framework under consideration. Hence, we shall refer to such a test as the *alternate student-t test* and briefly discuss its properties in the following.

Under the condition that $\eta$ is fixed, the null distribution of $\hat{\tau}_1$ is straightforward to derive – it is once again the student-t distribution with $n-k-1$ degrees of freedom – and the critical value $\tau_\alpha$ may now be found by solving the equation $1-\Psi_{n-k-1}(\tau_\alpha)=\alpha$, which yields

$$\tau_\alpha = \Psi_{n-k-1}^{-1}(1-\alpha) < t_\alpha$$

Therefore, the *alternate* test is *operational* and its properties are *tractable*. To that end, we invoke the commonly accepted approach to formal examination of the properties of a hypothesis testing procedure by focusing on the implied *power function*.

Recall that the *power* of a test is the *probability of rejecting a null hypothesis given that the alternative hypothesis is true*. The corresponding *power function* generates this probability for a particular value of the parameter permitted under the alternative hypothesis. Moreover, a test is *unbiased* if and only if its power is greater than or equal to the significance level for *all* parameter values under the alternative hypothesis, *consistent* if and only if the power converges to one as the sample size approaches infinity, and it is called *uniformly most powerful (UMP)* if and only if it exhibits greater power than *all other tests* (or within a class of tests) for each value of the parameter under the alternative hypothesis (for further discussions regarding this terminology, see for example (Poirier, 1995, pp. 351-373), (Judge, Hill, Griffiths, Lütkepohl, & Lee, 1988, pp. 92-109), (Greene, 2003, pp. 892-894), and (Goldberger, 1991, pp. 217-220)).

To the extent that we are interested in the properties of the *alternate* test defined above in comparison with the standard student-t test, we restate for convenience some well known properties of the student-t test. First note that its power function may be written as

$$\Pi_1(\mu) = \Psi_{n-k-1}\left(-t_\alpha; \mu\sqrt{n}\right) + \Psi_{n-k-1}\left(-t_\alpha; -\mu\sqrt{n}\right) \qquad (14)$$

where $\Psi_{n-k-1}(\,\cdot\,;\xi)$ denotes the *cdf* of the *noncentral-t* distribution with *noncentrality* parameter $\xi$ and $\mu = \frac{\beta_1}{\sigma}\sqrt{\frac{1}{nq_{22}}}$ (hence, $\mu \neq 0$ iff $\beta_1 \neq 0$). It follows that the student-t test is

[4] *Note that for $\eta$ to be a "fixed probability," it must be set according to known quantities not involving the dependent variable $y$; thus, it may include $n$, $k$, or any function of the explanatory variables (i.e., $R^2_{1,k}$, $\det(X'X)$, etc.) since the CNR treats $x_1,\dots,x_k$ as nonstochastic.*

P1(a) *unbiased* (i.e. $\Pi_1(\mu) \geq \Pi_1(0)$ for all $\mu \neq 0$)

P1(b) *UMP* amongst all *unbiased* tests

P1(c) *consistent* (i.e. $\Pi_1(\mu) \to 1$ as $n \to \infty$ for all $\mu \neq 0$)

P1(d) *robust* to the normality of errors assumption in large samples

P1(e) *robust* to the normality of errors assumption in small samples to the extent that properties P1(a)-P1(c) "hold when the normality assumption is generalized to other members of the *spherical family* of distributions." (Poirier, 1995, p. 502)

P1(f) closely related to *confidence intervals* associated with the parameters of interest

Similar to (14), the power function of the *alternate* test may be expressed as

$$\Pi_2(\mu) = \eta\Psi_{n-k-1}\left(-\tau_\alpha; \mu\sqrt{n}\right) + (1-\eta)\Psi_{n-k-1}\left(-\tau_\alpha; -\mu\sqrt{n}\right) \qquad (15)$$

For ease of demonstration, the comparison of the functions in (14) and (15) is illustrated graphically in Figure 1. Panels (A) and (B) consider the functions with $n-k-1=100$ degrees of freedom whereas Panels (C) and (D) represent the analogous plots with $n-k-1=250$ degrees of freedom. Moreover, Panels (A) and (C) depict the power function of the alternate test $\Pi_2(\mu)$ for several values of $0 < \eta \leq 0.5$ while the mirror cases with $0.5 \leq \eta < 1$ are captured in Panels (B) and (D). A quick examination of Figure 1 readily yields the following properties of the alternate test:

P2(a) when $\eta = .5$, it is *unbiased* and *uniformly less powerful* than the standard student-t test

P2(b) when $\eta = 0$, it is *biased*, more powerful for $\mu > 0$ and less powerful for $\mu < 0$

P2(c) when $0 < \eta < .5$, it is *biased*, more powerful for $0 < \mu < \mu^*$ and less powerful elsewhere[5]

P2(d) when $\eta = 1$, it is *biased*, more powerful for $\mu < 0$ and less powerful for $\mu > 0$

P2(e) when $.5 < \eta < 1$, it is *biased*, more powerful for $-\mu^* < \mu < 0$ and less powerful elsewhere

P2(f) when $0 \leq \eta \leq .5$, as $n$ increases $\Pi_2(\mu) \to \eta$ for $\mu < 0$ and $\Pi_2(\mu) \to 1-\eta$ for $\mu > 0$

P2(g) when $.5 \leq \eta \leq 1$, as $n$ increases $\Pi_2(\mu) \to 1-\eta$ for $\mu < 0$ and $\Pi_2(\mu) \to \eta$ for $\mu > 0$

While it is immediately evident by P2(a) that setting $\eta = 0.5$ always generates an *inferior* test in comparison with the standard student-t test, one interesting implication of P2(b) and P2(d) above is further apparent: if we reconsider the significance test

[5] *$\mu^*$ is the value obtained as the solution to the equation*

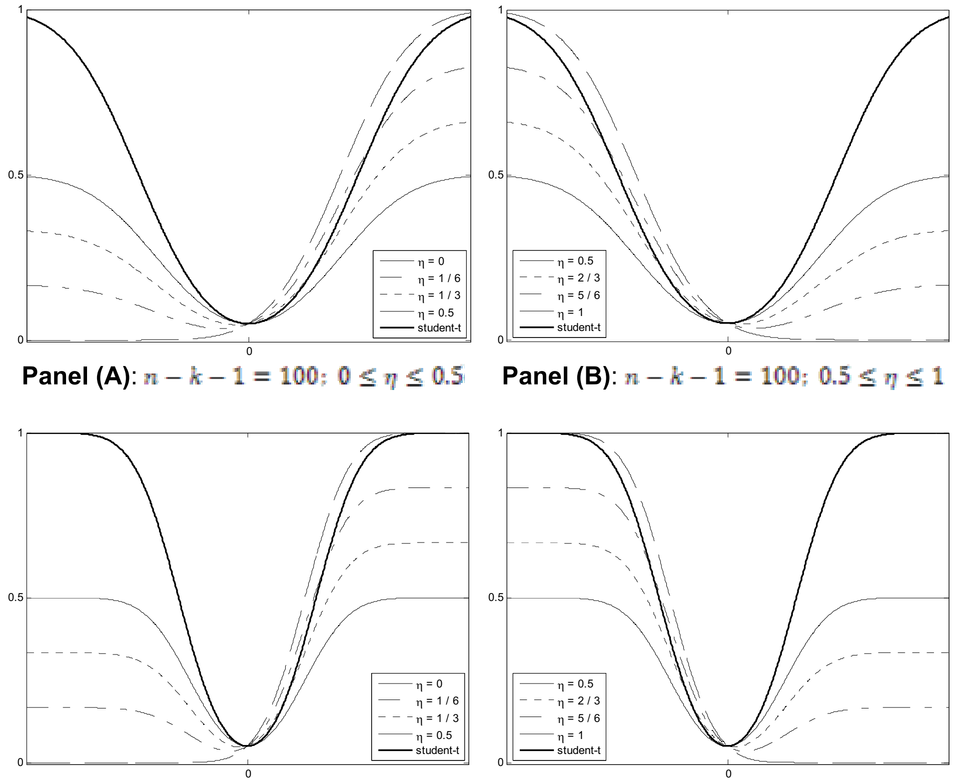
$$\begin{aligned}\eta\Psi_{n-k-1}\left(-\tau_\alpha; \mu^*\sqrt{n}\right) + (1-\eta)\Psi_{n-k-1}\left(-\tau_\alpha; -\mu^*\sqrt{n}\right) \\ = \Psi_{n-k-1}\left(-\tau_\alpha; \mu^*\sqrt{n}\right) + \Psi_{n-k-1}\left(-\tau_\alpha; -\mu^*\sqrt{n}\right)\end{aligned}$$

with $\mu^* > 0$ for a given value of $\eta < 0.5$. Note that if $\mu^*$ solves the equation above for $\eta < 0.5$, then $-\mu^*$ is the corresponding solution for $1-\eta$.

with a *one-sided* alternative hypothesis (e.g. either $H_A : \beta_1 > 0$ or $H_A : \beta_1 < 0$), then the *alternate* test is *unbiased* and *UMP* for appropriate assignments of $\eta = 0$ and $\eta = 1$. Note, however, that for $\eta = 0$, the alternate student-t test is identical to the *one-tailed* standard student-t hypothesis test with $H_A : \beta_1 > 0$ (analogously, $\eta = 1$ corresponds to the one-tailed test with $H_A : \beta_1 < 0$). In this sense, an interesting interpretation of the alternate test is that of a weighted combination of the one-tailed and two-tailed student-t tests.

**Figure 1**

**Plots of the Power Functions**



**Panel (A)**: $n - k - 1 = 100;\ 0 \le \eta \le 0.5$ **Panel (B)**: $n - k - 1 = 100;\ 0.5 \le \eta \le 1$

**Panel (C)**: $n - k - 1 = 250;\ 0 \le \eta \le 0.5$ **Panel (D)**: $n - k - 1 = 250;\ 0.5 \le \eta \le 1$

Regarding the alternate testing procedure as a balance between the one-tailed and two-tailed student-t tests, the question we seek to answer is when might such balancing procedure be useful? That is, when is it desirable to set $\eta$ strictly greater than zero or strictly less than one? The answer lies in the *a priori* knowledge of the researcher regarding $\mu$. To that end, suppose this knowledge may be classified as follows:

C1. Any value of $-\infty < \mu < \infty$ is equally likely (i.e. no prior information)
C2. Any value of $\mu \geq 0$ is equally likely; $\mu < 0$ is impossible
C3. Any value of $-\infty < \mu < \infty$ is possible; values $0 < \mu < \mu^*$ are *heavily favoured*
C4. Any value of $\mu \leq 0$ is equally likely; $\mu > 0$ is impossible
C5. Any value of $-\infty < \mu < \infty$ is possible; values $-\mu^* < \mu < 0$ are *heavily favoured*

Now, consider first the choice of hypothesis test under C1. In this case, we should clearly prefer a test which is *unbiased* over the entire support of $\mu$, and since the two-tailed student-t test is UMP in this class of tests, it is the obvious choice. In the case described by C2, on the other hand, we should readily sacrifice any testing power over the region $\mu < 0$ in favour of increased power over the region $\mu > 0$, and hence, the *most efficient* test under such prior beliefs is the one-tailed student-t test (or equivalently the alternate test with $\eta = 0$).

If our belief is according to C3, we focus our preference on a test which exhibits reasonably high power over the region $0 < \mu < \mu^*$, but not entirely void of power for negative values of $\mu$ close to zero. In particular, consider the *alternate* procedure with $\eta < 0.5$ corresponding to $\mu^*$ in comparison with the one-tailed and two-tailed student-t tests in application to this setting. While the one-tailed student-t test is more powerful over $0 < \mu < \mu^*$ than the relevant alternate test, we might consider that this difference in power is insufficient to justify the extensive power loss characterizing the one-tailed test in reference to the region $\mu < 0$. Conversely, while the two-tailed test is more powerful over $\mu < 0$, the alternate test dominates in power over the focal region of interest $0 < \mu < \mu^*$. Thus, depending on *how* likely we believe $\mu$ to be in the region $0 < \mu < \mu^*$ relative to $\mu < 0$, there may quite possibly exist strong justification for preferring the alternate test with $0 < \eta < 0.5$ to either the two-tailed or one-tailed standard student-t tests. We emphasize, however, that by P1(c) and P2(f)-P2(g), the region $0 < \mu < \mu^*$ where the alternate test might be more powerful than the student-t test shrinks as $n$ grows. The latter conclusion, therefore, is admittedly more applicable to small sample cases relative to large samples.

The purpose of the above analysis is to underscore two significant comments with respect to a more general and realistic search for a testing procedure targeting an improvement upon the standard student-t test. First, any worthwhile competing test will *inevitably* exhibit sharp trade-offs in comparison with the student-t test. That is, a test which improves upon the student-t test in terms of some properties P1(a)-P1(f) without sacrificing others does not exist, even in special cases (so much is obvious from P1(a) and P1(b) alone). On the other hand, it is difficult to imagine a trade-off that is *always* acceptable *in general*, as this would clearly imply that whatever property is being sacrificed under the said trade-off is *generally* unimportant. Accordingly, the best outcome of the search that one can realistically hope for is a test that generates a favourable trade-off in certain special cases. Yet, this may be of immense practical value to a researcher! We emphasis, therefore, that while it is not in the least the intention of this comment to discourage any future search for improved testing procedures, we simply stress that it is absolutely imperative to present a rigorous formal treatment in comparing the discussed properties between the proposed test and the student-t test.

As a second point of interest, we turn our focus specifically on improvements in terms of testing power. In this sense, we claim that in close resemblance to the alternate testing procedure analysis above, the relationship of the "special cases" to prior information defining favourable application of such a test will extend to the general search for competing test procedures. In particular, if there exists a practically sensible test which is preferred power wise to the standard student-t test in a certain case, such a special case will arise according to the researcher's prior belief regarding possible values of $\mu$ over certain regions, and more specifically, a belief relating the degree to which $\mu$ is more likely to fall in a certain region relative to other regions. The testing procedure search under these conditions, therefore, is closely related to the *probabilistic* a priori belief regarding the parameter values.

The focus on improvement in testing power, of course, is propelled by the fundamental motivation being cast in terms of uncovering a test procedure that improves statistical inference in presence of multicollinearity. In that sense, we return to the conventional view that the main adverse effect of collinearity on inference is attributed to the decrease in testing power. It remains, therefore, to resolve this motivating foundation with the motivation laid out in (Pavelescu, 2009). In fact, a careful interpretation of the latter text reveals that the two motivations are closely aligned, although stated in slightly different terminology.[6] The discussion of Section 3, in particular, merits a closer examination.

Section 3 of (Pavelescu, 2009) is dedicated to defining algebraic conditions under which a student-t statistic decreases in absolute magnitude when an additional explanatory variable is introduced into the linear regression. Stated formally, denote the student-t statistic defined in (6) as $t_{1,k}$, thereby relating that it is the statistic obtained in a significance test of $\beta_1$ using the estimates derived from the $k$-factorial specification of (1). The analogous $t_{1,k+1}$ is then the student-t statistic obtained from the $(k+1)$-factorial regression. The section, consequently, concentrates on the conditions (in terms of collinearity) under which $|t_{1,k+1}| \leq |t_{1,k}|$. The author argues that this comparison is important since holding $n$ fixed, increasing $k$ necessarily

[6] *Note that certain comments in (Pavelescu, 2009), such as*

*"...taking into account only the absolute values of the Student test, it may lead to the situation where 'statistical illusions' are considered as very good estimations for the proposed model" (p. 66)*

*may be misinterpreted as implying that the search for an improved testing procedure is predicated on a desire to reduce to the risk of erroneously rejecting the null hypothesis. Insofar as this risk corresponds precisely to the definition of Type I error, we assume that such an interpretation is not the author's intention and do not pursue it further. That is, the ability of the researcher to explicitly control the probability of Type I error occurrence lies at the very foundation of the hypothesis testing paradigm and merits no further explanation.*

reduces the degrees of freedom in the estimation, and therefore, leads to the increase in critical value $t_{1,k+1} > t_{1,k}$.

Observe that underlying these comparisons of $\hat{t}_{1,k+1}$, $\hat{t}_{1,k}$, $t_{1,k+1}$, and $t_{1,k}$ is a concern regarding the loss in testing power of the significance test related to the $(k+1)$-factorial regression in comparison with the test related to the $k$-factorial regression. Additionally, the conclusion drawn in Section 3 is that *any* reduction in the magnitude of $|\hat{t}_{1,k+1}|$ when compared to $|\hat{t}_{1,k}|$ is *inefficient*, and to that end, the reasonable interpretation of the *corrected* student-t statistic must be that it systematically counteracts the algebraic forces that reduce $|\hat{t}_{1,k+1}|$, thereby mitigating this inefficient decrease. We note, however, that casting the statistical power discussion in terms of such a test statistic magnitude comparison introduces a certain degree of confusion.

In terms of motivation, assigning *inefficiency* to the relative decrease of $|\hat{t}_{1,k+1}|$, in general, is largely misleading. To see this, consider first the case where the $(k+1)$-factorial regression is the *true* model (i.e. $\beta_{k+1} \neq 0$), and denote the OLS parameter estimators of $\beta_j$ generated by regressing $y$ on $x_1, \dots, x_{k+1}$ as $\hat{\gamma}_j$. In this case, employing the *short* model in (1) yields inference that is subject to the famous *omitted variable bias*. In particular, the sampling distributions derived in (4)-(5) are no longer correct. To appropriately characterize the resulting sampling distributions, let $\hat{\delta}_{1,k+1}$ be the OLS slope estimator in the auxillary regression of $x_{k+1}$ on the *residual* $(x_1 - \hat{x}_1)$, which is in-turn obtained from regressing $x_1$ on the remaining explanatory variables $x_2, \dots, x_k$. Then the correct sampling distributions of $\hat{\sigma}^2$ and $\hat{\beta}_1$ (as defined in (2)-(3)) are

$$(n-k-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-k-1;\ .5\beta_{k+1}^2/\mathrm{var}(\hat{\gamma}_{k+1})) \tag{16}$$

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1 + \delta_{1,k+1}\beta_{k+1}, \sigma^2 q_{22}\right) \tag{17}$$

where $\chi^2(\nu;\ \xi)$ represents the *noncentral chi-square* distribution with $\nu$ degrees of freedom and *noncentrality* parameter $\xi$. Consequently, the *bias* of the estimator $\hat{\beta}_1$ is $\delta_{1,j}\beta_{k+1}$, while the *bias* of the estimator $\hat{\sigma}^2$ is $\frac{\beta_{k+1}^2}{\mathrm{var}(\hat{\gamma}_{k+1})} \cdot \frac{\sigma^2}{n-k-1}$. It is clear, therefore, that unless $\beta_{k+1} = 0$, the test statistic $\hat{t}_{1,k}$ *does not* follow a student-t distribution. In fact, its sampling distribution in this case depends on the nuisance parameters $\beta_{k+1}$, $\sigma^2$, and consequently, attempting to employ this test statistic in practice gives rise to precisely the same operational difficulties relevant to the *corrected* student-t statistic discussed in the previous section (it is certainly illogical to compare its value to $t_{1,k} = \Psi_{n-k-1}^{-1}(1-\alpha/2)$ or any other value found in the student-t tables).

The student-t statistic $\hat{t}_{1,k+1}$ obtained from the full regression, on the other hand, *does* follow the student-t distribution with $n-k-2$ degrees of freedom and is readily usable in carrying out statistical inference. Regardless of how one defines *inefficiency*, therefore, working with the correctly specified $(k+1)$-factorial model and the resulting $\hat{t}_{1,k+1}$ is certainly *desirable*. In general, however, since $\hat{t}_{1,k+1}$ and $\hat{t}_{1,k}$ are characterized by different sampling distributions (especially with $\hat{t}_{1,k}$ following an intractable distribution), there is admittedly little sense in comparing their magnitudes in the presence of omitted variable bias.

The case where the discussion of (Pavelescu, 2009), Section 3, is more relevant, perhaps, is when we are sure that $\beta_{k+1} = 0$. Then, the *short* model with $k$ explanatory variables is the *true* model, while the $(k+1)$-factorial model is said to be *overspecified*. Overspecification, however, does not introduce bias in the parameter estimates; rather, including the excess variable $x_{k+1}$ only inflates their standard errors (i.e. $\hat{\gamma}_0, \dots, \hat{\gamma}_k$ are still unbiased, but are no longer *minimum-variance* estimators). Therefore, although both $\hat{t}_{1,k+1}$ and $\hat{t}_{1,k}$ follow student-t distributions (with $n-k-2$ and $n-k-1$ degrees of freedom, respectively), significance testing with $\hat{t}_{1,k+1}$ is necessarily less powerful than the corresponding test with $\hat{t}_{1,k}$, and in this sense, inference derived with the $(k+1)$-factorial regression is always less efficient than inference derived with the $k$-factorial regression. On other hand, if we are sure that $\beta_{k+1} = 0$, there obviously exists no realistic scenario under which we would employ the $(k+1)$-factorial model.

Following this logic, a quest for a *more efficient test statistic* must therefore resemble a search for a test statistic that reflects the properties of $\hat{t}_{1,k}$ when $\beta_{k+1} = 0$ and $\hat{t}_{1,k+1}$ when $\beta_{k+1} \neq 0$, in the realistic case when $\beta_{k+1}$ is unknown. Alternatively stated, we seek a test statistic that systematically identifies the case $\beta_{k+1} = 0$ *more efficiently* than $\hat{t}_{1,k+1}$. An econometrician in pursuit of such a statistic, however, must readily admit that this is a tantamount (with inclination towards insurmountable) task by simply observing that (regarding $\beta_{k+1}$ as unknown) $\hat{t}_{1,k+1}$ is already constructed from the *minimum variance* estimator of $\beta_{k+1}$, namely $\hat{\gamma}_{k+1}$. In what sense, therefore, may one expect an alternative statistic to identify $\beta_{k+1} = 0$ more accurately than $\hat{t}_{1,k+1}$?

In contrast to the preceding discussion, it appears advisable to maintain focus in motivating a search for an improved testing procedure on the effect of collinearity within the realm of a *correctly specified model*. As discussed in Section 1 of the present text, the epicentre of this effect lies in the variance of the individual parameter estimate under consideration. To that end, however, it should be noted that the effect of collinearity on such estimates is significant *only* relative to the sample size. In fact, if the joint explanatory vector $(x_{i1} \cdots x_{ik})$ is independent across observations $i$, the effect (on the variance of an individual parameter estimate) of an *increase* in the sample size by one observation is equivalent to the effect of a *decrease* in the auxiliary coefficient of determination $R^2_{j,k}$ by $\frac{1}{n}(1 - R^2_{j,k})$. Intuitively, both collinearity and sample size may be viewed as two very similar factors that determine the variability in the sample, which is the primary source of information offered by the data for statistical inference. Hence, the extent of either effect (high collinearity or low sample size) on individual parameter inference must be interpreted accordingly.

In bringing the discussion on the severity of collinearity to a close, we refer to the point of view offered by the esteemed econometrician Arthur S. Goldberger (Goldberger, 1991, p. 252):

> *To say that "standard errors are inflated by multicollinearity" is to suggest that they are artificially, or spuriously, large. But in fact they are appropriately large: the coefficient estimates actually would vary a lot from sample to sample. This may be regrettable but it is not spurious.*

Note that from a purely classical perspective that obstinately refutes all prior information in statistical inference, this claim is undisputable. That is, one certainly cannot commit inferential exclusivity to a set of data, and upon receiving vague inference from that data, dismiss this vagueness on the grounds that the data is "poorly conditioned." Within the bonds of data exclusivity, one simply has no way of judging to what extent is a confidence interval "unreasonably" wide, since there exists no basis for comparison in succinctly defining "unreasonably." Hence, it is not "the relevance of the estimated parameters" that is diminished by collinearity (Pavelescu, 2009), rather only the ability of the data to identify their relevance, and without an ulterior source of information this deficiency is incircumventable; different techniques of manipulating the data can only exploit more efficiently (or less efficiently) the variability already present, but will never induce more variability.

On the other hand, it is entirely reasonable to gauge the degree to which "standard errors are inflated" by admitting prior information because the prior belief provides exactly the basis for comparison lacking in a purely classical paradigm. More specifically, one may justifiably claim that a confidence interval is "too wide" if it extends into regions where *a priori* the researcher assigns a low degree of probability in the sense that the information offered by the data regarding the parameter of interest contradicts the prior information. In addition, the extent of this "contradiction" may be sensibly measured when the prior information is formalized in a probabilistic manner. It is in this sense that prior information provides an additional instrument, which in the presence of collinearity provides the crucial supplement to variability lacking in the data. This, in turn, reaffirms our previous claim that hypothesis test improvements in presence of collinearity must be closely linked to the use of probabilistic prior information in statistical inference.

Insofar as an investigator would be willing to admit such prior beliefs in conducting inference, however, it would be natural for her to consider the Bayesian framework as an alternative to the classical hypothesis testing centred methods altogether. Whereas the example at the beginning of this section involving the alternate testing procedure demonstrates the awkwardness associated with incorporating prior information in classical hypothesis testing, Bayesian methods are well known to be the most efficient way of systematically combining prior information with the data in generating robust statistical inference (for introductory Bayesian texts, see (Koop, 2003) and (Gelman, Carlin, Stern, & Rubin, 2003)). To that end, we conclude with a simple demonstration of how prior information may be employed in alleviating adverse effects of collinearity within the Bayesian linear regression framework (Koop, 2003, pp. 15-85), (Gelman, Carlin, Stern, & Rubin, 2003, pp. 351-385), (Poirier, 1995, pp. 524-580).

## 4. Bayesian Example

In concept, Bayesian inference differs fundamentally from classical inference in the following sense: the focus of Bayesian inference is on what the parameter is *most likely to be*, whereas the most common concern of classical inference is on what the parameter is *definitely not*. Nevertheless, there are strong practical parallels between the two approaches. For example, given a particular significance level $\alpha$ the $(1-\alpha)\%$ *posterior probability interval* (commonly constructed as the *highest posterior density (HPD)* interval) bears a close resemblance to the $(1-\alpha)\%$ *confidence interval* for either individual parameters or a combination of parameters, while the *mode* of the posterior distribution is comparable to the parameter *estimate* generated by classical techniques. More importantly, as the sample size increases, both the posterior modes and posterior probability intervals converge to the corresponding *Maximum Likelihood* estimates and confidence intervals (Poirier, 1995, pp. 306-307).

Note that the latter fact reflects exactly the previously outlined intuition regarding the effect of collinearity relative to sample size. Insofar as the effect of collinearity is most apparent in smaller samples and diminishes proportionally as $n$ increases, it is crucial that whatever instrument is adapted to offset the effects of collinearity in smaller samples reduces in relative importance as the sample size grows. Employing prior information through Bayesian techniques achieves just that: prior information is most influential on the posterior distribution, and hence most effective in combating collinearity, when $n$ is small, while this influence is proportionately reduced as $n$ increases and vanishes altogether as $n \to \infty$.

We illustrate the Bayesian approach in this context through a simple simulation example based on some well-known results of the Bayesian linear regression. Accordingly, suppose the model of interest is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2) \tag{18}$$

and the *true* parameter values are

$$\beta_0 = 15, \qquad \beta_1 = 1, \qquad \beta_2 = -4, \qquad \sigma^2 = 2$$

Using a *pseudo-random* number generator, we simulate three data samples of $\{y_i, x_{i1}, x_{i2}\}$ for $n = 100$, $n = 1{,}000$, and $n = 10{,}000$, respectively, and compare the inference one would obtain under the Bayesian framework to that of the classical framework. Since our primary interest lies in the influence of collinearity on statistical inference, the simulated data is generated to yield a relatively high sample correlation between the explanatory variables while each pair $\{x_{i1}, x_{i2}\}$ is sampled independently. Moreover, the experiment is designed such that the correlation between $x_1$ and $x_2$ emits a stronger effect on the precision of the estimates of $\beta_1$ relative to $\beta_2$. Summary statistics of the simulated datasets are reported in Table 1.

Now, consider how an econometrician might approach the task of estimating this model aware only of the descriptive properties of the data and operating under the assumption that the linear model is *correctly specified* as given in (18).

Table 1

Simulation Data Summary Statistics

| | $n$ | average | standard deviation | correlation $y$ | correlation $x_1$ | correlation $x_2$ |
|---|---|---|---|---|---|---|
| $y$ | 100 | −70.424 | 2.411 | 1.000 | −0.733 | −0.840 |
| $x_1$ | 100 | 5.033 | 0.258 | | 1.000 | 0.891 |
| $x_2$ | 100 | 22.596 | 0.554 | | | 1.000 |
| $y$ | 1,000 | −69.976 | 2.421 | 1.000 | −0.703 | −0.815 |
| $x_1$ | 1,000 | 4.987 | 0.246 | | 1.000 | 0.893 |
| $x_2$ | 1,000 | 22.487 | 0.559 | | | 1.000 |
| $y$ | 10,000 | −70.007 | 2.470 | 1.000 | −0.702 | −0.811 |
| $x_1$ | 10,000 | 5.003 | 0.247 | | 1.000 | 0.893 |
| $x_2$ | 10,000 | 22.504 | 0.551 | | | 1.000 |

Assume further that the econometrician is in possession of the smallest sample ($n = 100$) and is concerned that the *strength* of collinearity relative to this sample size may lead to uninterestingly vague inference regarding her primary parameters of interest $\beta_1$ and $\beta_2$. On the other hand, her theoretical training endows her with some key intuition regarding the values of these parameters. She summarizes her beliefs as follows:

1. centred at $\beta_1 = 2$, $\beta_2 = -2$
2. symmetric (i.e. $\beta_1 = 0$ is just as likely as $\beta_1 = 4$, etc.)
3. highly unlikely that $\beta_1 > 5$ or $\beta_2 < -5$

These beliefs may be formalized in terms of *prior probability distributions* regarding $\beta_1$ and $\beta_2$. Consequently, we shall proceed with a general form of the prior distribution given by

$$\beta_j \mid \sigma^2 \overset{\text{nid}}{\sim} \mathcal{N}\left(m_{j+1}, \sigma^2 v_{j+1,j+1}^2\right), \quad \sigma^2 \sim \mathcal{IG}\left(\frac{g}{2}, \frac{h}{2}\right) \tag{19}$$

where $\mathcal{IG}(\cdot)$ denotes the *inverse gamma distribution* (for example, see (Gelman, Carlin, Stern, & Rubin, 2003, pp. 573-577)). It can be shown that the implied *marginal* distribution of $\beta_j$ is

$$\frac{\beta_j - m_{j+1}}{v_{j+1,j+1}\sqrt{h/g}} \sim t(g) \tag{20}$$

and therefore, all three prior beliefs described above may be accommodated in (20) by appropriately setting the parameters $m_{j+1}$ and $v_{j+1,j+1}^2$. Specifically, let $m_2 = 2$, $m_3 = -2$, $v_{2,2}^2 = v_{3,3}^2 = \frac{3g}{h}/\Psi_g^{-1}(c)$. This ensures that the modes of the distributions for $\beta_1$ and $\beta_2$ are $2$ and $-2$, respectively, while $\Pr(\beta_1 > 5) = \Pr(\beta_2 \leq -5) = c$, where $c$ may be set to any reasonably small value, (e.g. the ensuing results are based on

$c = 0.01$). The symmetry condition is, of course, automatically satisfied since the student-t distribution is naturally symmetric.

Note that Bayesian methods require that prior distributions be properly specified for *all* parameters. Since, the researcher is neither particularly interested in $\beta_0$ and $\sigma^2$, nor does she posses very specific beliefs regarding their values, she may specify $g$, $h$, $m_1$, and $v_{1,1}^2$ in such way that results in *mildly-informative* prior distributions for $\beta_0$ and $\sigma^2$. Such mild beliefs, for example, are sufficiently represented with the following values: $g = 2$, $h = 2$, $m_1 = 0$, $v_{1,1}^2 = 100$. It is worthwhile to observe that using this complete prior specification, the researcher may derive the relevant distribution of $\mu$, identify the implied region where $\mu$ is most probable and attempt to design a hypothesis test that dominates in power over that region at the expense of being inferior over the less probable regions. In relevance to the discussion of the previous section, we mention that while the implied distribution of $\mu$ is complicated form both theoretical and practical aspects in general, the present beliefs regarding

$\beta_1$ and $\sigma^2$ imply that $\Pr(0 < \mu \leq 0.35) \approx 0.75$. Thus, for example, the *alternate* hypothesis test previously outlined is more powerful over the region $(0, 0.35]$ relative to the student-t test for $\eta = 0.035$ at $n = 100$ (for larger $n$, the corresponding $\eta \approx 0$).

On the other hand, if we choose to proceed with the Bayesian inference, we focus on *updating* our prior belief by the observed data sample. This, in turn, requires the construction of the *likelihood* function, which is operationally expressed as the distribution of the dependent variable conditional on the parameters:

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n) \tag{21}$$

where $\boldsymbol{I}_n$ denotes the $n \times n$ *identity matrix*. Using (19) and (21), the *joint posterior* distribution is obtained through Bayes' Rule as

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\beta}, \sigma^2)\, p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2)}{p(\boldsymbol{y})} \tag{22}$$

and contains all information necessary to carry out statistical inference on the parameters. While exact analytic expressions for posterior distributions are, in general, intractable (most often, posterior inference is based on *simulating* from the posterior distribution), the Bayesian linear regression model yields fairly simple and practically straightforward posteriors.

Consequently, define the following notation: let $\boldsymbol{m} = (m_1\ m_2\ m_3)'$, $\boldsymbol{V} = \mathrm{diag}\{(v_{1,1}^2\ v_{2,2}^2\ v_{3,3}^2)\}$ (i.e. a $3 \times 3$ diagonal matrix), and

$$\begin{aligned}
\hat{\boldsymbol{V}} &= (\boldsymbol{V}^{-1} + \boldsymbol{X}'\boldsymbol{X})^{-1} \\
\hat{\boldsymbol{m}} &= \bar{\boldsymbol{V}}(\boldsymbol{V}^{-1}\boldsymbol{m} + \boldsymbol{X}'\boldsymbol{y}) \\
\hat{h} &= h + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{m}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{m}}) + (\hat{\boldsymbol{m}} - \boldsymbol{m})'\boldsymbol{V}^{-1}(\hat{\boldsymbol{m}} - \boldsymbol{m}) \\
\hat{g} &= g + n
\end{aligned} \tag{23}$$

The joint posterior distribution of all model parameters in our case is then given by

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y} \sim \mathcal{N}\left(\hat{\boldsymbol{m}}, \sigma^2 \hat{\boldsymbol{V}}\right), \quad \sigma^2 \mid \boldsymbol{y} \sim \mathcal{IG}\left(\frac{\hat{g}}{2}, \frac{\hat{h}}{2}\right) \tag{24}$$

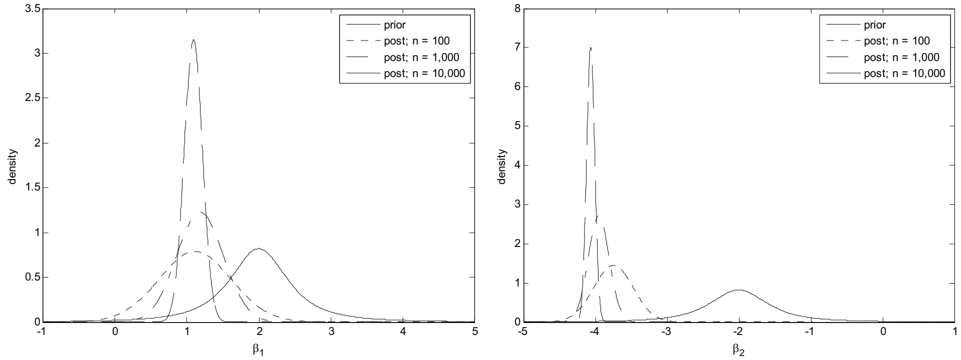
whereas the *marginal* posterior distributions of interest are obtained as

$$\frac{\beta_j - \hat{m}_{j+1}}{\hat{v}_{j+1,j+1}\sqrt{\hat{h}/\hat{g}}} \mid \boldsymbol{y} \sim t(\hat{g}) \tag{25}$$

The marginal posterior distributions for each of the three cases (of varying sample size) under examination are plotted along with the corresponding prior distributions for the parameters $\beta_1$ and $\beta_2$ in Figure 2, Panels (A) and (B), respectively. More specifically, Panel (A) illustrates the evolution of the posterior distribution of $\beta_1$ from the prior as the sample size grows while Panel (B) depicts the analogous phenomenon for $\beta_2$. The intuition regarding the influence of prior information on posterior inference as $n$ increases is immediately evident. In both cases, with each increasing sample size, the posterior distribution *collapses* around the mode, which in turn, converges to the true parameter value.

**Figure 2**

**Prior and posterior distributions for $\beta_1$ and $\beta_2$**



**Panel (A)**: $\beta_1$ **Panel (B)**: $\beta_2$

Observe, however, that the *collapsing* effect is distinctly slower for $\beta_1$ in comparison to $\beta_2$. This is precisely a reflection of the influence of collinearity, which is by design more influential in the posterior distribution of $\beta_1$than that of $\beta_2$. In fact, a closer examination of Panel (A) reveals that the posterior of $\beta_1$ for $n = 100$ is not noticeably less *dispersed* relative to its prior distribution, but rather only exhibits a shift in *location* towards the true value. A more illuminating interpretation of the latter may be formulated as follows: the posterior distribution of $\beta_1$ at $n = 100$ reflects a *joint* effort on the part of the prior information and the data whereby the information from the data is incorporated into more accurately centring the posterior while the prior maintains the dispersion contained by substituting for the lack of certainty projected by highly collinear data with *a priori* information. As a result, even with a relatively low sample size (i.e. relative to the degree of correlation in the explanatory variables), posterior inference regarding $\beta_1$ is sufficiently informative.

The important trade-off is, of course, that this gain in precision at $n = 100$ is strongly reliant on the prior beliefs, and hence, accentuates the importance of introducing prior information cautiously and in a manner that is convincingly justifiable. On the other hand, as the sample size grows and the information projected by the data gains in vigour, the need for the prior to contain the posterior precision diminishes and its role in determining the shape of the posterior distribution is marginalized. This is clearly reflected in Panel (A), by the progressive reduction of the posterior dispersion at $n = 1{,}000$ and $n = 10{,}000$ where the abundance of available observations overcomes the small sample deficiencies resulting from collinearity.

Table 2 and Table 3 further reinforce this intuition in a numerical comparison of Bayesian and classical inference that would be conventionally employed in interpreting the results for each of the three sample size levels.

**Table 2**

**Bayesian/Posterior Inference**

|  | $n$ | mean / median / mode | standard deviation | 95% Probability Interval (HPD) |  | zero outside interval |
|---|---|---|---|---|---|---|
|  |  |  |  | lower bound | upper bound |  |
| $\beta_1$ | 100 | 1.113 | 0.510 | 0.111 | 2.115 | yes |
|  | 1,000 | 1.182 | 0.326 | 0.544 | 1.821 | yes |
|  | 10,000 | 1.089 | 0.127 | 0.840 | 1.337 | yes |
| $\beta_2$ | 100 | −3.747 | 0.279 | −4.294 | −3.200 | yes |
|  | 1,000 | −3.951 | 0.147 | −4.240 | −3.662 | yes |
|  | 10,000 | −4.064 | 0.057 | −4.175 | −3.952 | yes |

**Table 3**

**Classical/Frequentist Inference**

|  | $n$ | estimate | standard deviation | 95% Confidence Interval |  | significance test |  |
|---|---|---|---|---|---|---|---|
|  |  |  |  | lower bound | upper bound | $\hat{t}$ | significantly different from zero |
| $\beta_1$ | 100 | 0.743 | 1.135 | −1.510 | 2.996 | 0.655 | no |
|  | 1,000 | 1.201 | 0.399 | 0.419 | 1.983 | 3.013 | yes |
|  | 10,000 | 1.090 | 0.130 | 0.836 | 1.345 | 8.392 | yes |
| $\beta_2$ | 100 | −3.965 | 0.528 | −5.012 | −2.917 | −7.511 | yes |
|  | 1,000 | −4.001 | 0.175 | −4.346 | −3.657 | −22.802 | yes |
|  | 10,000 | −4.069 | 0.058 | −4.183 | −3.955 | −69.964 | yes |

A quick overview of Table 3, which summarizes typical classical quantities of interest, reveals the diminishing effect of collinearity in increasing $n$: for both parameters $\beta_1$ and $\beta_2$, as $n$ increases regression estimates converge to the true values, standard errors decrease, confidence intervals shrink, and significance test statistics grow (in absolute value). Additionally, this phenomenon is accelerated for quantities related to $\beta_2$, in evident parallel with the influence of prior information on posterior distributions, and is likewise explained by the more prominent influence of collinearity on the precision of $\beta_1$ estimates.

In this sense, the fact that the analogous posterior quantities detailed in Table 2 converge to their classical counterparts is unsurprising. In fact, at $n = 10{,}000$ the posterior modes of $\beta_1$ and $\beta_2$ (which for the student-t distribution are equivalent to the respective posterior means and medians) are nearly identical to the OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. Similarly, posterior standard deviations are approximately equivalent to the standard errors, as are the $95\%$ posterior probability intervals to the $95\%$ confidence intervals.

Where the posterior inference differs most significantly from classical inference, however, is in terms of the parameter $\beta_1$ for $n = 100$. Here, it is worthwhile to note that the $95\%$ classical confidence interval extends over negative values of $\beta_1$. Indeed, the limited sample is not informative enough to identify $\beta_1$ as statistically significantly different from zero at the $5\%$ significance level in the classical context (this is equivalently verified by the corresponding significance test failing to reject the null hypothesis $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$). On the Bayesian side, Table 2 illustrates that with a sample of $n = 100$ observations, the *lower bound* of the $95\%$ posterior probability interval is notably *greater than* the hypothesized null value $\beta_1 = 0$. In an analogous statement of significance, therefore, our Bayesian inference allows us to confidently proclaim $\beta_1$ as statistically significantly different from zero, given our prior beliefs.[7]

Incidentally, the *corrected* student-t statistic of (Pavelescu, 2009) for $\beta_1$ is *negative* at all levels of $n$ in our example. This is easily verified by observing from Table 1 that $r_{y,x_1} < 0$ in all three cases while according to Table 3, $t_1 > 0$. In consequence, heeding the instructions prescribed in (Pavelescu, 2009, pp. 66-67), we should fail to "validate" this regression, which is *correctly specified by experimental design*, regardless of the available sample size.

[7] In fact, the Bayesian paradigm defines a formal methodology of Bayesian Hypothesis Testing which is based on Bayesian Posterior Odds and is generally unrelated in terms of inference to the HPD interval approach demonstrated here; for more details see (Poirier, 1995, pp. 376-392, 540-551). However, the technical and conceptual complexities involved in a satisfactory discussion of posterior odds are beyond the scope of our purpose. We only mention here that in the simplified setting of our example, and particularly insofar as our focus is on comparing posterior inference to classical inference, the HPD interval approach is sufficiently appropriate.

## References

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003), "*Bayesian Data Analysis"* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Goldberger, A. S. (1991), "*A Course in Econometrics",* Cambridge: Harvard University Press.

Greene, W. H. (2003), "*Econometric Analysis",* (5th ed.). New Jersey: Prentice Hall.

Johnston, J., & DiNardo, J. (1997). *Econometric Methods* (4th ed.). New York: McGraw-Hill.

Judge, G. G., Hill, R. C., Griffiths, W. E., Lütkepohl, H., & Lee, T.-C. (1988), "*Introduction to the Theory and Practice of Econometrics",* (2nd ed.). New York: John Wiley & Sons.

Koop, G. (2003), "*Bayesian Econometrics",* Chichester: John Wiley & Sons, Ltd.

Pavelescu, F.-M. (2009), "A Review of Student Test Properties in Condition of Multifactorial Linear Regression", *Romanian Journal of Economic Forecasting , 10* (1): 63-75.

Poirier, D. J. (1995),. "*Intermediate Statistics and Econometrics: A Comparative Approach",* Cambridge: The MIT Press.