THE TRUE R^2 AND THE TRUTH ABOUT R^2

Nicolas Christou (nchristo@stat.ucla.edu)
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095

KEY WORDS: Population coefficient of determination; Sample coefficient of determination; Noncentral F distribution; Regression analysis.

Abstract

In this paper our goal is to explain the distribution of the sample coefficient of determination in the simple regression case. We do this by using its relationship to the noncentral F distribution. But first we introduce a new term, the true coefficient of determination. In a simulation study it is feasible to know the true coefficient of determination because the variance of the error term is known. The usefulness of the true coefficient of determination is in the built of relationships with predetermined strength. It answers the question: How much error should we add? The answer depends on how strong we want the association in the simple regression model to be. Once we determine this we can compute the noncentrality parameter and explain the distribution of the sample coefficient of determination. It is a simple way of explaining the distribution of the sample coefficient of determination and it is interesting at least from the educational point of view.

1 Introduction

We discuss the distribution of the sample coefficient of determination $\hat{R}^2$ and its relationship to the noncentral F distribution. The calculation of $\hat{R}^2$ is a basic result in the method of regression analysis. However in teaching regression analysis, we almost never get into the details of its distribution. Its relationship to the noncentral F distribution has an interesting result that we discuss here. The noncentral F

distribution which is taught in a mathematical statistics course in connection with the power of the F test in analysis of variance gives us the tools to explain easily the distribution of $\hat{R}^2$. We need to define a new term - the true (or population) coefficient of determination (R^2). In computing the regression of Y on x using the simple regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, we should always calculate the sample $\hat{R}^2$. The true R^2 is simply the measurement of the true linear association between Y and x when the variance of the error term is known. Franklin (1992) used simulated data to estimate the parameters of the above model when the error term follows normal distribution. Lee (1971) discussed the sampling distribution of the multiple correlation coefficient. In this paper we present a simple way to explain the distribution of $\hat{R}^2$ through the noncentral F distribution, and we introduce the idea of the true R^2. The true R^2 enables us to compute the true strength of the association between Y and x when different amounts of error are added into the model. First we define the true R^2 and how it is computed and we describe how the simulations were performed. We then present the results and at the end we explain the distribution of $\hat{R}^2$.

2 Computing the true R-squared

We call true R^2 the population R-squared and we define it as the squared correlation coefficient between y and x. That is

$$R^2 = \frac{cov^2(y,x)}{\sigma_x^2 \sigma_y^2}$$

where $cov(y,x)$ is the covariance between y and x. Because $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ it is easy to see that $cov(y,x) = \beta_1 \sigma_x^2$ and $var(y) = \beta_1^2 \sigma_x^2 + \sigma_\epsilon^2$. Substituting these two expression into R^2 we get

$$R^2 = \frac{(\beta_1 \sigma_x^2)^2}{\sigma_x^2(\beta_1^2 \sigma_x^2 + \sigma_\epsilon^2)} = \frac{\beta_1^2 \sigma_x^2}{\beta_1^2 \sigma_x^2 + \sigma_\epsilon^2} \tag{1}$$

The previous result is true when X is random and also X and ϵ are independent. The sample squared coefficient of correlation is define as $\hat{R}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2}$ where

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}(x_i-\bar{x})y_i \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1 x_i+\epsilon_i) \\ &= \beta_1 s_x^2 + \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})\epsilon_i = \beta_1 s_x^2 + s_{x\epsilon} \end{aligned}$$

and

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\Big(\beta_1(x_i-\bar{x})^2 + (\epsilon_i-\bar{\epsilon})\Big)^2 = \beta_1^2 s_x^2 + 2\beta_1 s_{x\epsilon} + s_\epsilon^2$$

Therefore

$$\hat{R}^2 = \frac{(\beta_1 s_x^2 + s_{x\epsilon})^2}{s_x^2(\beta_1^2 s_x^2 + 2\beta_1 s_{x\epsilon} + s_\epsilon^2)}$$

This $\hat{R}^2$ converges to the true R^2 because s_x^2 converges to σ_x^2 and $s_{x\epsilon} = 0$.

Therefore in a simulation study we can choose how much R^2 we want and this will tell us how much error we must add to the signal. From (1) above we find that

$$\sigma_\epsilon^2 = \frac{\beta_1^2\sigma_x^2(1-R^2)}{R^2} \tag{2}$$

Clearly from the previous expression we can see that when we put $R^2 = 1$ there is no noise in the model ($\sigma_\epsilon = 0$). By doing this the student can see the impact of adding more error and how this affects the relationship between y and x. Even though x is random below we work conditionally on the x values. Therefore when the x values are given the distribution of the sample $\hat{R}^2$ follows the non-central F distribution as we will see later.

3 Simulations

The first step is to use a deterministic model $y_i = \beta_0 + \beta_1 x_i$. Suppose we choose to use $\beta_0 = 2, \beta_1 = 2$ and $x_i = 1 + 0.1i, i = 1, \cdots, 10$. The variance of these data

Model	True R^2 (%)	σ_ϵ^2
1	90	0.0407
2	80	0.0917
3	70	0.1571
4	60	0.2444
5	50	0.3667
6	40	0.5500
7	30	0.8556
8	20	1.4667
9	10	3.3000

Table 1: True R^2 and the variance of the error term.

.

is $\sigma_y^2 = \beta_1^2\sigma_x^2 = 2^2\sigma_x^2 = 0.36667$. We can see from the scatterplot on Figure 1 (a) that the points fall on a straight line. There is no surprise here that for this perfect relationship $R^2 = 100\%$. This is because $R^2 = (\frac{cov(x,y)}{\sigma_y\sigma_x})^2 = (\frac{\beta_1\sigma_x^2}{\beta_1\sigma_x\sigma_x})^2 = 1$. However, because in statistics we are interested in relationships that are not perfect let us add some error to the above model. The model becomes $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where we assume normal distribution for the error term. But how much error do we want to add? This depends on how strong we want the relationship between y and x to be. For example, if we want true $R^2 = 90\%$ then according to (2) above $\sigma_\epsilon^2 = 0.0407$. Similarly for other choices of the true R^2 we construct Table 1. Also on Figure 1 b-i the scatterplots of y on x are shown for true $R^2 = 0.90, (0.10), 0.20$. The purpose of these simulations is to show to students that when we generate data there is an inherent strength between the dependent and the independent variables. This strength is measured by the true R^2 as defined in (1). This true R^2 also exists in real data but there it cannot be computed because we do not know the true variance of the error term. After we generate the data we have only available the observed values of y and x. Using the method of least squares we will try to find the relationship that generated these data.

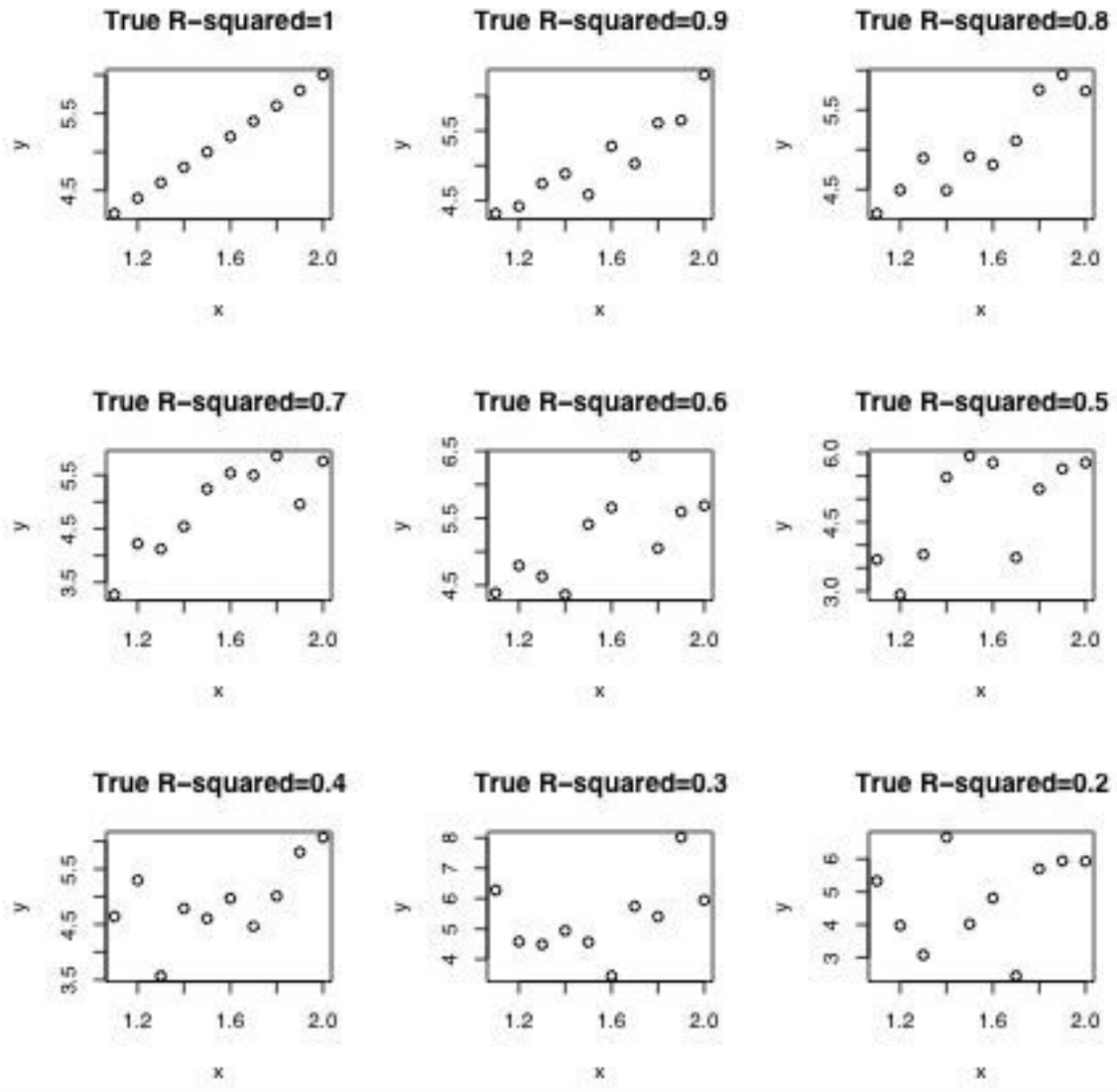

Figure 1: Simulated (a to i) scatterplots of y on x for true $R^2 = 20\% - 100\%$

4 Results

After the data generation we have data on y and x. Let us pretend now that we know nothing about the relationship that exists between the two variables. Using the method of least squares we obtain for each model (one for each case that corresponds to Table 1) estimates of the parameters. We generated 1000 different data sets for each case to obtain estimates on $\beta_0, \beta_1, \sigma^2$, and R^2 showing on Table 2. This $\hat{R}^2$ is the usual coefficient of determination. The mean, median, first and third quartiles of these estimates are presented in Table 2. An interesting comment can be made here that if the sample variance of the n generated error terms happens (unlikely though) to be equal to the true variance of the error term the usual R^2 will be exactly equal to the true R^2. In Table 2 we observe the following: The $\hat{R}^2$ always overestimates the true R^2 with increasing overestimation as the variance of the noise increases. This is not a surprise! Consider the extreme situation where $R^2 = 0$. Even in this situation $\hat{R}^2$ will never be zero (unless $\beta_1 = 0$). As we know always $0 < \hat{R}^2 < 1$. We also observe that $\hat{\sigma}^2_\epsilon$ underestimates the true σ^2_ϵ. In addition to the above results we constructed the frequency distributions for each estimate (using the 1000 values). The results are shown on Figures 2-5. As expected, the histogram of the 1000 values of $\hat{\sigma}^2_\epsilon$ resembles the χ^2 distribution (Figure 3). This is consistent with the theory that the estimate of σ^2_ϵ follows the χ^2 distribution with $n-2$ degrees of freedom. Also, from theory we know that the estimates of β_0 and β_1 follow the normal distribution. This can be seen on Figures 4 and 5. However, the most interesting distribution is that of the $\hat{R}^2$ (Figure 2). We observe that when there is little amount of noise in the model (this corresponds to high values of true R^2) the distribution of the $\hat{R}^2$ is left-skewed, while when there is large amount of noise (this corresponds to low values of the true R^2) the distribution of $\hat{R}^2$ is right-skewed. This can be explained intuitively first. Consider for example the case when the true R^2 is very high (say 90 %). Most of the time the value of $\hat{R}^2$ will be around 90%. We know that $\hat{R}^2$ is bounded by 0 and 1, so there is enough room to the left of 90% and this can explain

			$\hat{R}^2$				$\hat{\sigma}^2$			
Model	True R^2 (%)	σ_ϵ^2	Q_1	Median	Q_3	Mean	Q_1	Median	Q_3	Mean
1	90	0.0407	0.89	0.92	0.94	0.91	0.0255	0.0370	0.0526	0.0381
2	80	0.0917	0.75	0.83	0.88	0.81	0.0564	0.0822	0.1208	0.0854
3	70	0.1571	0.64	0.74	0.81	0.72	0.0999	0.1449	0.2017	0.1495
4	60	0.2444	0.51	0.65	0.75	0.62	0.1548	0.2253	0.3172	0.2293
5	50	0.3667	0.40	0.55	0.67	0.53	0.2383	0.3519	0.4844	0.3581
6	40	0.5500	0.27	0.45	0.61	0.44	0.3430	0.4946	0.7122	0.5188
7	30	0.8556	0.18	0.35	0.52	0.36	0.5454	0.7862	1.0795	0.7969
8	20	1.4667	0.08	0.24	0.39	0.26	0.9526	1.3428	1.8627	1.3988
9	10	3.3000	0.04	0.14	0.29	0.19	2.1404	3.0730	4.3193	3.1702

			$\hat{\beta}_0$				$\hat{\beta}_1$			
Model	True R^2 (%)	σ_ϵ^2	Q_1	Median	Q_3	Mean	Q_1	Median	Q_3	Mean
1	90	0.0407	1.791	2.024	2.236	2.018	1.841	1.985	2.134	1.989
2	80	0.0917	1.698	2.030	2.406	2.038	1.741	1.979	2.187	1.974
3	70	0.1571	1.556	1.987	2.419	1.996	1.725	2.014	2.283	2.004
4	60	0.2444	1.402	1.973	2.5889	2.004	1.630	2.040	2.369	2.000
5	50	0.3667	1.222	1.939	2.697	1.971	1.588	2.042	2.493	2.023
6	40	0.5500	1.081	1.987	2.841	1.967	1.465	2.008	2.560	2.019
7	30	0.8556	0.904	1.946	3.060	1.986	1.370	2.033	2.695	2.011
8	20	1.4667	0.663	1.993	3.456	2.047	1.037	2.014	2.811	1.966
9	10	3.3000	-0.221	1.879	4.066	1.884	0.695	2.117	3.394	2.074

Table 2: results

the left skewness. However, we will try more formally to explain this behavior of $\hat{R}^2$ in the next section.

5 Explaining the distribution of $\hat{R}^2$

One possible way to explain the distribution of $\hat{R}^2$ is through its density $f(\hat{R}^2)$. This density is a complicated function and we will avoid using it here (Lee 1971, 1972). A nice way to explain this distribution is by using the fact that the ratio

$$F = \frac{\hat{R}^2}{1-\hat{R}^2}(n-2) \tag{3}$$

under the alternative hypothesis that $\beta_1 \neq 0$ follows the noncentral F distribution with noncentrality parameter $\gamma^2 = \frac{(n-1)\sigma_X^2\beta_1^2}{\sigma_\epsilon^2}$ and degrees of freedom 1 for numerator and $n-2$ for denominator (Hogg and Craig 1998). Rearranging (3) we get $\hat{R}^2 = \frac{F}{F+n-2}$. Suppose now that we want to compute the probability that $\hat{R}^2$ is less than

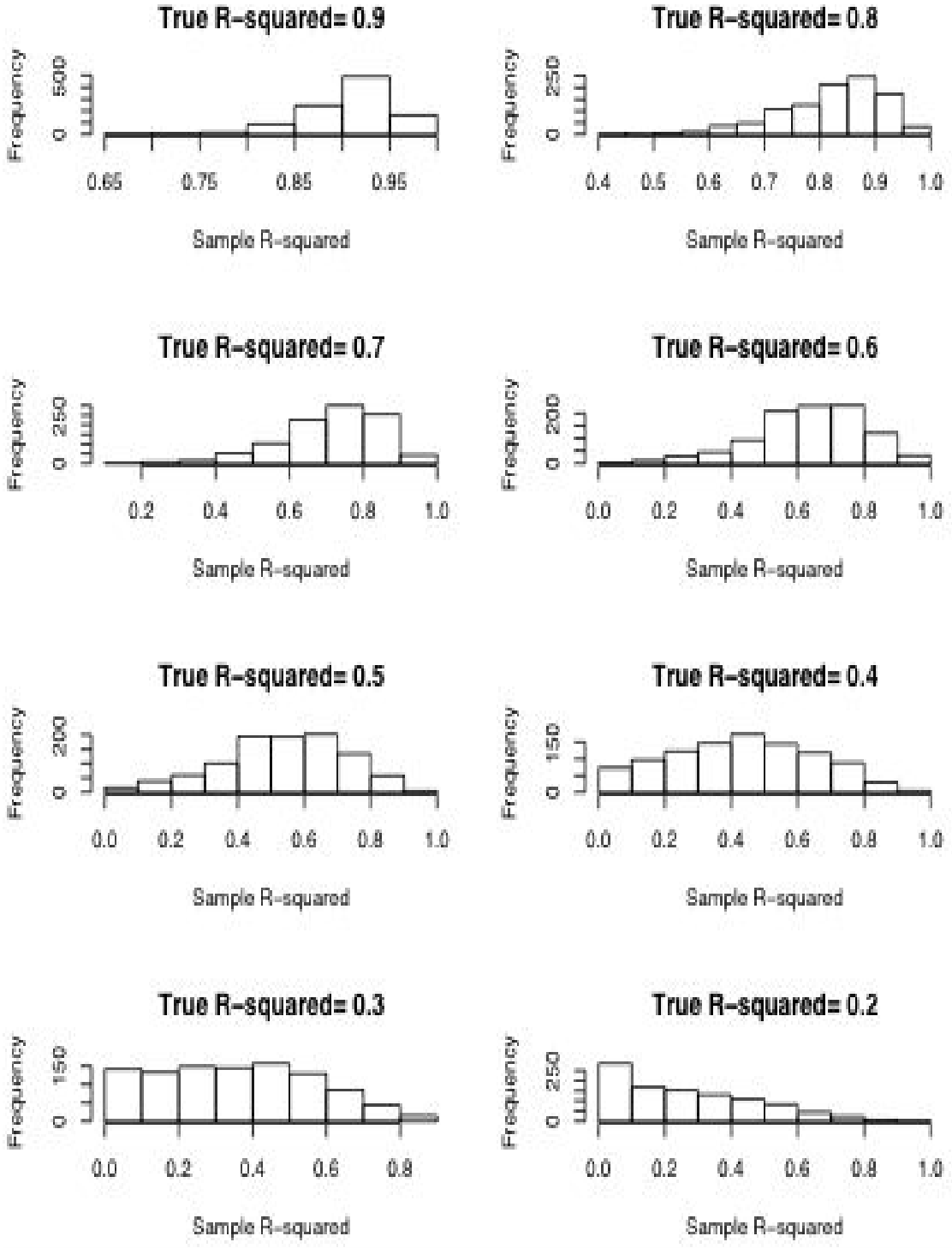

Figure 2: Frequency of $\hat{R}^2$ for true $R^2 = 20\% - 90\%$

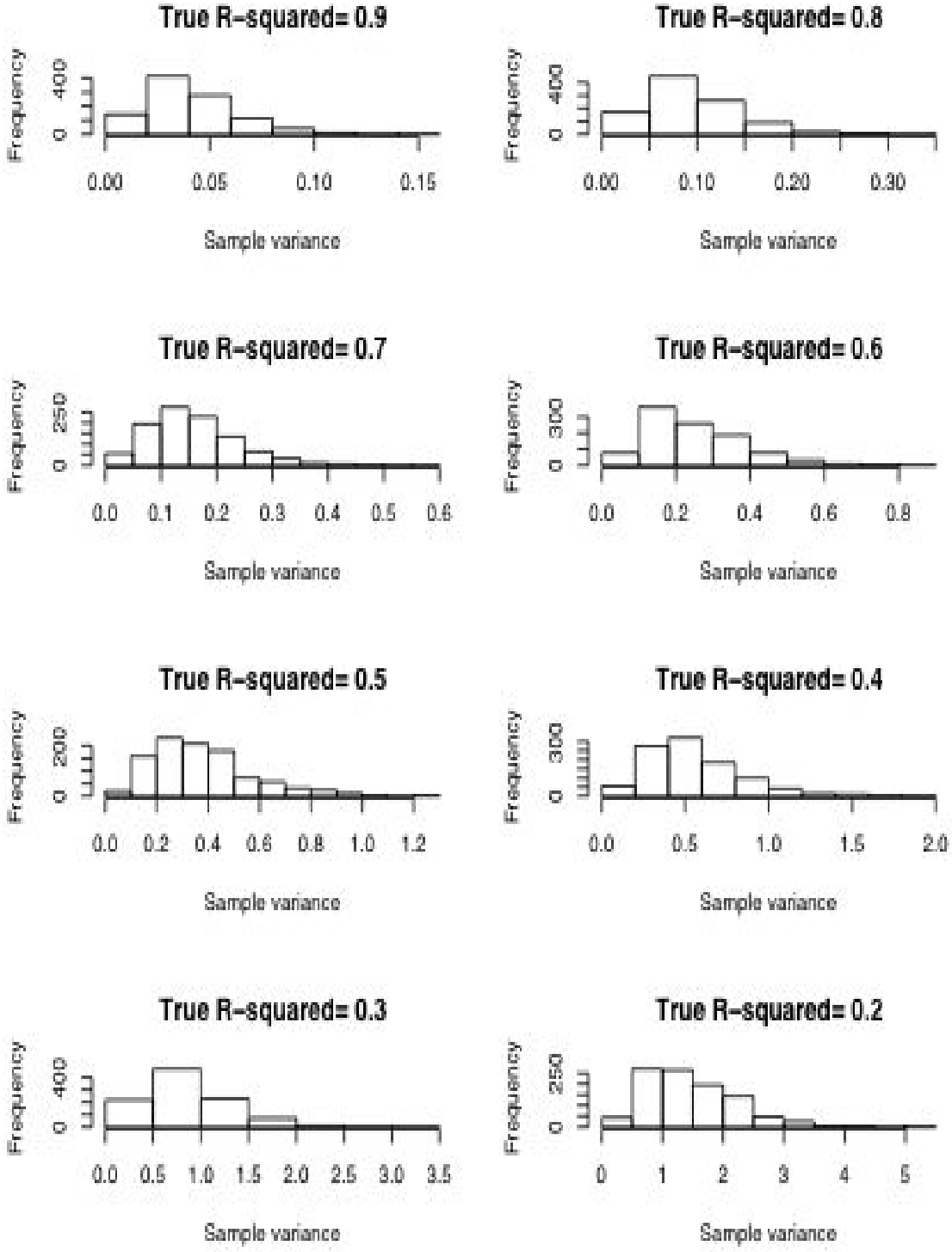

Figure 3: Frequency of $\hat{\sigma}_\epsilon^2$ for true $R^2 = 20\% - 90\%$

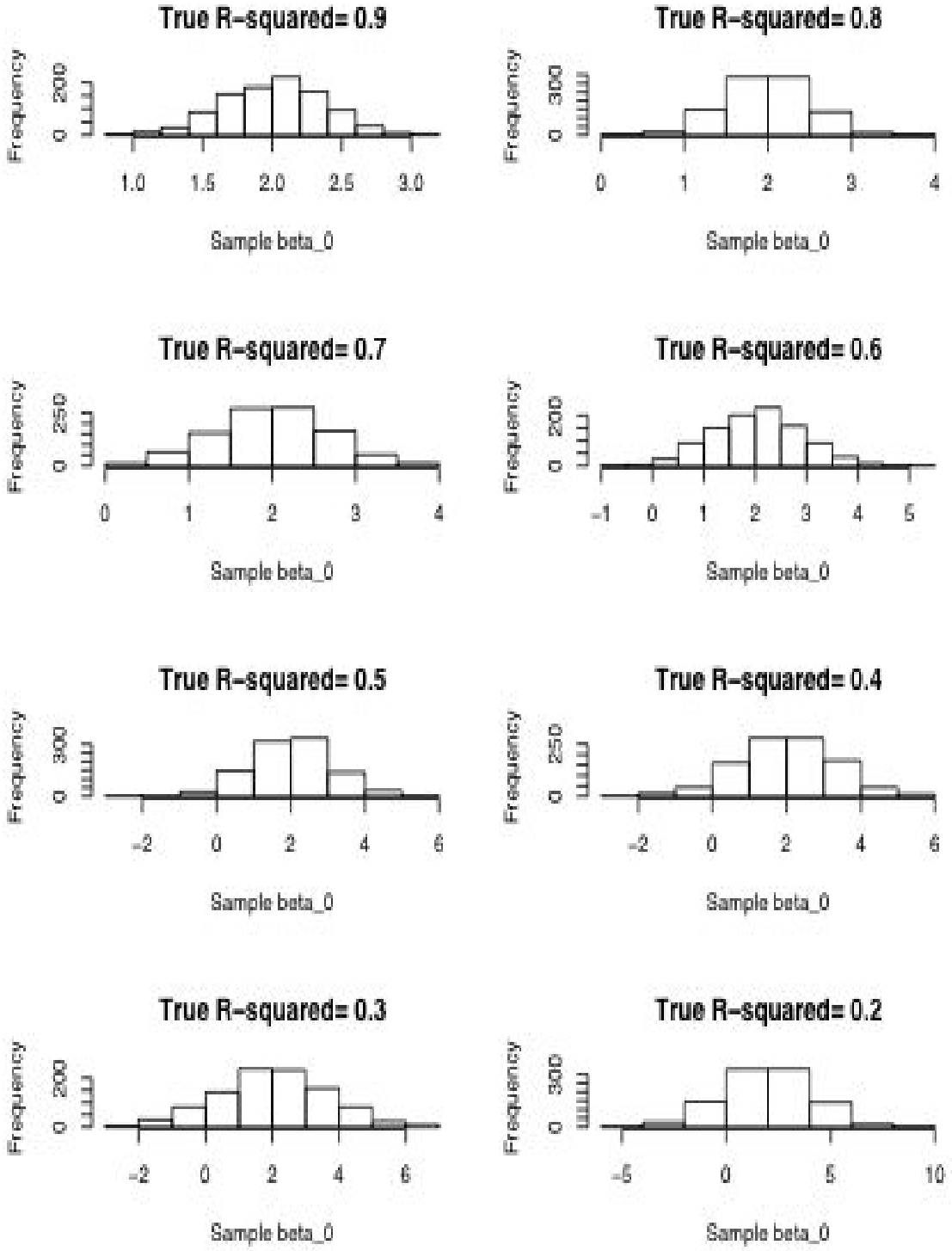

Figure 4: Frequency of $\hat{\beta}_0$ for true $R^2 = 20\% - 90\%$

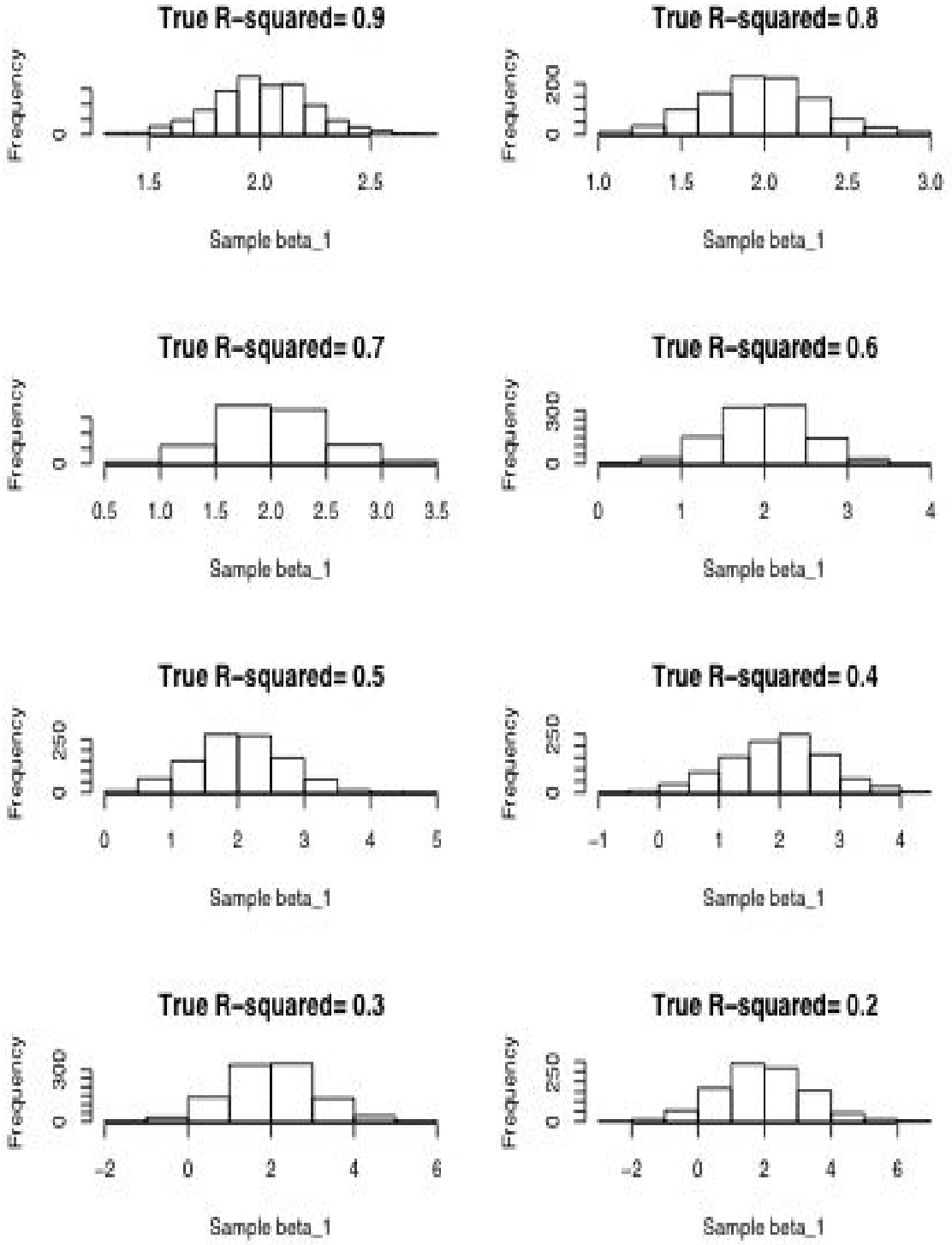

Figure 5: Frequency of $\hat{\beta}_1$ for true $R^2 = 20\% - 90\%$

		$P(a < \hat{R}^2 < b)$									
True R^2	γ^2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.00	.000	.627	.178	.094	.051	.028	.014	.006	.002	.000	.000
0.10	1.000	.435	.189	.135	.097	.066	.041	.022	.009	.002	.002
0.20	2.250	.275	.173	.156	.135	.108	.077	.047	.022	.006	.002
0.30	3.857	.151	.136	.151	.155	.144	.119	.083	.044	.014	.002
0.40	6.000	.067	.087	.121	.150	.166	.161	.131	.081	.030	.004
0.50	8.999	.021	.041	.074	.116	.160	.190	.187	.139	.061	.010
0.60	13.502	.004	.011	.029	.062	.114	.179	.231	.223	.126	.020
0.70	21.006	.000	.001	.005	.016	.045	.108	.212	.306	.254	.052
0.80	35.987	.000	.000	.000	.001	.004	.020	.085	.265	.456	.170
0.90	81.081	.000	.000	.000	.000	.000	.000	.001	.023	.334	.642

Table 3: Percentiles of the $\hat{R}^2$ distribution

ζ, where $0 < \zeta < 1$. We can do this using:

$$P(\hat{R}^2 < \zeta) = P\left(\frac{F}{F+n-2} < \zeta\right) = P\left(F < \frac{\zeta}{1-\zeta}(n-2)\right)$$

Therefore this probability can be computed simply by using a table of the noncentral F distribution. For each model we computed the noncentrality parameter and then we found $P(\hat{R}^2 < \zeta)$ for $\zeta = 0.1\ (0.1)\ 1.0$. Using these results we can construct Table 3 which gives the percentiles of the distribution of $\hat{R}^2$. The entries in the table are the exact probabilities that $\hat{R}^2$ is between two consecutive values of ζ. For example let us consider the case when the true $R^2 = 0.40$. The noncentrality parameter is computed to be $\gamma = 6$. Then the exact probability that $\hat{R}^2$ is between 10% and 20% is 0.087. Also, 20% is the $0.067 + 0.087 = 0.154$ or the 15.4_{th} percentile of the distribution of $\hat{R}^2$. If we now look at the frequency distribution of $\hat{R}^2$ when the true $R^2 = 0.40$ (Figure 2) we can find that about 95 cases out of 1000 or 0.095 have $\hat{R}^2$ between 10% and 20%. This compares very well with the exact probability. Similarly the other entries can be matched to the bars of the frequency distribution.

6 Conclusion

In this paper we explained the distribution of the coefficient of determination ($\hat{R}^2$) and how this is related to the noncentral F distribution. We usually never go into the details about this important statistic but here we present some characteristics of it. We see how it is related to the population R^2 (the true R^2) and conclude that it overestimates it. Under the null hypothesis that the slope is zero ($H_0 : \beta_1 = 0$) the distribution of $\hat{R}^2$ is related to that of the central F. Under the alternative hypothesis ($H_a : \beta_1 \neq 0$) it is related to the noncentral F distribution. Therefore we can use both of them to compute the power for different values of R^2. We also see that the shape of distribution of $\hat{R}^2$ varies depending on how much error is added to the model. Students must first understand what a strong relationship between two variables is. Starting with a deterministic model and showing that $R^2 = 1$ then we can create data that come from models with less R^2 by changing the variance of the error term. This proposed method of teaching regression analysis assumes that the students must be familiar with correlation, random errors, probabilistic and deterministic models. We would also agree with Mills (2002) that these methods of teaching regression analysis would probably be more useful for students in a more advanced regression course rather than an introductory course. A student can appreciate more what a simulation can offer if he is familiar first with some concepts of modeling. Putting this aside, we believe that the idea of the term true R^2, which is not mentioned in most (if not all) textbooks it is important to help students understand the difference between population and sample in the regression case.

References

[1] Franklin, L.A. (1992). Using simulation to study linear regression. The College Mathematics Journal, 23, 290-295.

[2] Hogg, R.V., Craig, A.T. (1998). Introduction to Mathematical Statistics. Macmillan, Fifth Edition.

[3] Lee, Yoong-Sin. (1971). Some results on the sampling distribution of the multiple correlation coefficient. Journal of the Royal Statistical Society, Series B, 33, 117-130.

[4] Lee, Yoong-Sin. (1972). Tables of the upper percentage points of the multiple correlation coefficient. Biometrika, 59, 175-189.

[5] Mills Jamie D. (2002). Using computer simulation methods to teach statistics: A literature review. Journal of Statistics Education.

[6] Sen, A., Srivastava, M. (1990). Regression Analysis: Theory, Methods, and Applications. Springer-Verlag.