# Too Big to Fail:
# Larger Samples and False Discoveries

Mingfeng Lin, Henry C. Lucas, Jr., and Galit Shmueli*

## Abstract

The Internet presents great opportunities for research about information technology, allowing IS researchers to collect very large and rich datasets. It is common to see research papers with tens or even hundreds of thousands of data points, especially when reading about electronic commerce. Large samples are better than smaller samples in that they provide greater statistical power and produce more precise estimates. However, statistical inference using p-values does not scale up to large samples and often leads to erroneous conclusions. We find evidence of an over-reliance on p-values in large sample IS studies in top IS journals and conferences. In this commentary, we focus on interpreting effects of individual independent variables on a dependent variable in regression-type models. We discuss how p-values become deflated with a large sample and illustrate this deflation in analyzing data from over 340,000 digital camera auctions on eBay. The commentary recommends that IS researchers be more conservative in interpreting statistical significance in large sample studies, and instead, interpret results in terms of practical significance. In particular, we suggest that authors of large-sample IS studies report and discuss confidence intervals for independent variables of interest rather than coefficient signs and p-values. We also suggest taking advantage of a large dataset for examining how coefficients and p-values change as sample size increases, and for estimating models on multiple subsamples to further test robustness.

**Keywords:** large samples, p-value, statistical significance, practical significance, effect size, confidence intervals, Monte-Carlo

---

* Mingfeng Lin (mingfeng@eller.arizona.edu) is at the Department of MIS, Eller College of Management, University of Arizona. Henry C. Lucas (hlucas@rhsmith.umd.edu) and Galit Shmueli (gshmueli@rhsmith.umd.edu) are at the Department of Decision, Operations and Information Technologies, R.H. Smith School of Business, University of Maryland, College Park.

# Too Big To Fail:
# Larger Samples and False Discoveries

## Introduction and Purpose

The Internet and the advent of electronic commerce have provided information systems researchers with a wealth of data about markets and their participants. In the early days of IS research, a large sample size was over 100; today papers routinely report samples with tens or even hundreds of thousands of data points. While results from larger samples are usually considered better, is there a point at which a large sample size can lead to misleading results? The purpose of this research commentary is to point out potential difficulties in interpreting p-values from models that are based on very large samples and to propose steps to reduce the chance that models estimated from large samples lead to unwarranted conclusions. In particular we:

- Find that 47% of recent papers with sample sizes over 10,000 in the two leading IS journals and 57% of large sample papers in two leading IS conferences rely almost exclusively on low p-values to support their hypotheses.
- Recommend that researchers ignore statistical significance for results from large samples with coefficients having very low p-values.
- Advocate reporting and interpreting effect sizes and practical significance instead of focusing on p-values.
- Suggest basing effect size in directional hypotheses on the lower/upper bound of the appropriate confidence interval.
- If required to report statistical significance for individual variables, do so using a Monte Carlo simulation of smaller sample sizes to plot a distribution of *p*-values.
- Discuss Monte Carlo simulation, which allows the researcher to take advantage of large samples to estimate a model for multiple subsamples; the resulting distributions complement confidence intervals by enabling one to assess the robustness of the findings and to obtain confidence intervals for complex models where assumptions are likely to be violated.

## Background

Advances in technology have brought us the ability to collect, transfer, and store large datasets. Thanks to this, more and more empirical studies published in the information systems and related fields now rely on very large samples. Some samples include tens of thousands of observations: For example, Pavlou and Dimoka (2006) use "over 10,000 publicly available feedback text comments… in eBay"; Overby and Jap (2009) use "108,333 used vehicles offered in the wholesale automotive market"; Forman et al. (2008) "collected data on … [175,714] reviews from Amazon"; and finally, Goldfarb and Lu (2006) report "For our analysis, we have … 784,882 [portal visits]". Yet many papers still report, and more importantly, rely on, p-values for interpretation. Estimated linear regression models based on 500,000 observations are becoming common in the literature. However, p-values associated with coefficients from such models are very likely to be essentially 0, and this statistical measure becomes of little value.

Key questions arise that are both methodological and practical. From a methodological point of view, why is it that the statistical testing machinery "breaks down" when the sample is large? What exactly is the cause of the inability to scale up to large samples? Without understanding how sample size impacts p-values, how can one rely on statistical testing? How does a reader, including reviewers and editors, judge whether a small p-value in any sample is meaningful? In other words, can a sample be *too large*? We refer the readers to Appendix A for an explanation of how and why p values quickly approach 0 as sample sizes increase. [1]

[1] Most large sample IS research involves estimating regression models and testing the statistical significance of the coefficients of specific independent variables. It is important to remember that all statistical inference is based on accepting or rejecting a null hypothesis of no effect. In the regression case, the researcher usually has a directional hypothesis, for example, that the coefficient of a particular independent variable is positive and greater than 0.

## Large Sample Studies in IS: Literature Review

At a recent seminar at one of our universities, a researcher presented a paper with nearly 10,000 observations and discussed the regression results solely on the basis of statistical significance; there was no mention of effect sizes. This approach appears to be common practice in IS, as can be seen in recent papers in the top IS journals. We reviewed all papers published in the last five years in *MIS Quarterly, Information Systems Research* and *Management Science* along with abstracts from the *Workshop on Information Systems and Economics* and symposia on *Statistical Challenges in Electronic Commerce Research* to see to what extent IS researchers recognized the issues in analyzing large samples.[2] Table 1 reports the results. We find that 50% of recent papers with sample sizes over 10,000 in the two leading IS journals, *MIS Quarterly* and *Information Systems Research,* and 48% of large sample papers in two IS conferences (WISE and SCECR) rely almost exclusively on low p-values to support their hypotheses. Table 1 summarizes our literature review. It is interesting to note that there are an equal number of large-sample IS papers in *Management Science* as the two leading IS journals, and that almost all of the papers in this more general publication report practical significance.

---

The p-value of the coefficient is the probability that the researcher would have observed this coefficient or one more extreme if the null hypothesis is really true. See Appendix A for a more detailed discussion.

[2] Only abstracts are available for the two conference series so we have more confidence in the results for the two journals.

**Table 1: Large sample papers (n>10,000) in leading IS journals and conferences 2005-2010, broken down by method of reaching conclusions (practical significance, mostly p-values and coefficient sign, or solely p-values and coefficient sign)**

| | Conclusions rely on practical significance | Conclusions rely mostly on signs and p-values | Conclusions rely solely on signs and p-values | Total |
|---|---|---|---|---|
| **MISQ, ISR** | 2 | 1 | 6 | 9 |
| **Mgt Science** | 8 | 0 | 2 | 10 |
| **WISE (abstracts)** | 21 | - | 29 | 50 |
| **SCECR (abstracts)** | 12 | - | 15 | 27 |
| **Totals** | 44 (45%) | 4 (4%) | 51 (53%) | **97** |

Our conclusion is that information systems research that is based on large samples may be producing misleading results by over- relying on p-values to interpret findings. The purpose of this article is mainly to raise awareness to this critical issue. We also suggest ways to mitigate the deflating p-value problem, providing greater confidence in the results of large-sample IS research.

## Recommendations

In most of the information systems studies with large samples that we reviewed, the predominant method of analysis was to hypothesize about the direction of individual variable parameters (betas) in regression models. In most cases, the authors relied on finding a statistically significant coefficient in the hypothesized direction to confirm their propositions. We propose that large sample researchers be more conservative about interpreting statistical significance in large sample studies, and instead present and interpret effect sizes. What is the impact of a change in the independent variable on the dependent variable, all else held constant? It is frequently the case with large samples that a coefficient has a very low p-value, but also an extremely small effect size so that it has essentially no practical significance. The issue here is that of "too much power", where the large

sample “magnifies” even miniscule, unimportant effects.

Instead of discussing p-values and statistical significance, we recommend that for statistically significant coefficients, researchers focus on point estimates and their confidence intervals, for example, stating that the corresponding variable’s parameter (beta) would fall between *a* and *b* “95% of the time” or “with 95% certainty.” A conservative strategy for directional hypotheses (e.g., $H_1$: $\beta>0$ or $H_1$:$\beta<0$) is to report effect sizes based on the lower or upper bound of the confidence interval, depending on the hypothesized direction.

If the researcher is required to present statistical significance, we recommend supplementing the reporting of practical significance using a Monte Carlo approach to show the behavior of the coefficient and p-values as a function of sample size. This can be done by repeatedly drawing subsamples of increasing size from the large sample and estimating the researcher’s model. Then plot the resulting coefficient and p-value distributions for each variable of interest as a function of sample size to show how statistical significance compares to the desired level of significance, say 5% or 1%. Editors and reviewers of large sample papers can also request that researchers report such graphs when they have concerns of inference using large samples.

In addition to changing the method of interpreting and reporting results of large-sample studies, we also recommend taking advantage of the large sample for evaluating model assumptions on which the coefficient interpretations rely. This can be done by employing Monte Carlo to compute the empirical distribution of the coefficient, based on multiple subsamples from the large sample.

For statistically *in*significant coefficients, the researcher should ask the question “why are the results not significant?” Statistically insignificant results are especially surprising with very large samples, where one usually finds that most coefficients have low associated p-values. “Why?” is a particularly

interesting question when a regression coefficient is significant in the *opposite direction* from what the researcher predicted.[3]

In the next section, we present an example of the impact of a large sample on p-values for the coefficients of a model of factors affecting the sales price of digital cameras on eBay. This example illustrates and motivates our recommendations.

## Example: Research on eBay Auctions

Many IS researchers believe that a large sample size is always better in terms of discovering new effects by way of statistical inference, and in particular by low p-values. When a sample size is extremely large, very modest results – in terms of effect size –become statistically significant, often with p-value<.01 or p-value<.001. Since we are accustomed to looking for statistically significant results by examining the p-values, the casual reader may draw an unwarranted conclusion about the practical significance of a finding when there is little support from the data.

For the purpose of illustration, we draw on the work of Lucking-Reiley, Bryan, Prasad and Reeves (2007) (Henceforth referred to as LBPR). The original dataset used in LBPR includes eBay auction data on coins (collected using a web crawler) over a 30-day period during July and August of 1999. The results presented in the paper use a smaller dataset of US Indian Head pennies minted between 1859 and 1909, where only single items are listed in auctions, and the condition of the coin is known. There are 461 such auctions. The authors further strengthen their dataset using the book value of each coin. While the coins are interesting, in our replication we focus on some of their results and study how

[3] Statistical packages typically report a p-value for a regression coefficient based on a two-tailed test. If the researcher has predicted a direction, and the result is in that direction, then the correct p-value is half of the two-tailed value reported. If, however, the result is significant in the opposite direction of the prediction, the p value is really 1- (half of the reported p-value), which will be highly insignificant.

statistical significance changes with the sample size, using a different and much larger dataset from eBay. Specifically, we would like to replicate the following results from LBPR:

1. Higher minimum bids are associated with higher final auction price.
2. The presence of a reserve price is associated with higher sales price.
3. The longer the duration of auctions, the higher the price.
4. The more favorable the seller feedback, the higher the price.

Our intention is *not* to evaluate or criticize LBPR, since the sample used in their model is quite modest (400 or so). While our replication could shed light on the robustness of their results, our primary interest is adopting their model specification and using that specification to highlight how the statistical significance of coefficients changes when a much larger sample is available.

## Digital Camera Sales Data

Our new dataset contains all eBay auctions of digital cameras between August 2007 and January 2008. These data were purchased directly from eBay. Consistent with LBPR, we only study the auctions that sell one single item (instead of multiunit auctions). To be consistent with LBPR, we further remove store fixed price auctions, personal offers, and pure fixed price auctions.

The remaining auctions are the *Chinese* style listing (according to eBay's terminology) with one single item, which is also the most popular, comprising 84.39% of all listings in our dataset. The total dataset has 552,609 auctions (including those not resulting in sales, as consistent with LBPR). Among these, **341,136** auctions do not have a buy-it-now option and are pure auctions (bidders cannot end

auctions early by using the buy-it-now option[4]). This set of auctions constitutes the *overall* sample that we will be referring to in the paper.

It should be noted that while our empirical setting is not identical to the original paper, we are not actually interested in the sale of coins. While the earlier paper incorporates book value, a variable external to eBay, for digital cameras it is hard to justify a book value. We do, however, have a few additional characteristics regarding the digital cameras, including their condition (used/new/refurbished), camera type (point-and-shoot; SLR etc.), product line, brand and so on, which should be able to control for a "fair market value" of each product listed in this category.

Another difference between our dataset and the one used in LBPR is that they collected the number of positive, negative and neutral ratings, while we only observe the feedback score of a seller at the time of the auction This score is constructed from the number of positive and negative feedbacks the seller has received. While this reputation statistic is not as detailed, the feedback score should at least partially account for reputational effects. Table 2 shows the primary variables used in our model.

To test for the robustness of their findings, LBPR used four different model specifications. For illustrative purposes, however, we focus on one of their simplest specifications:

[4] $$\ln Price = \beta_0 + \beta_1 * \ln(minimumBid) + \beta_2 * reserve + \beta_3 * \ln(sellerFeedback) + \beta_4(Duration) + \beta_5(\text{controls}) + \varepsilon$$

It should be emphasized again that we are not attempting to replicate or test the results of LBPR, but rather, we are using this model as an example to demonstrate the sensitivity of the p-value when the sample size increases.

---

[4] It should be noted that while 211,473 auctions have the buy-it-now option, only 50,806 auctions are actually ended that way. To be cautious, however, we simply focus on the 341,136 auctions that do not allow buy-it-now from the beginning of listings.

**Table 2: Variable descriptions and summary statistics**

| Variables | Descriptions | Mean | Standard Deviation |
|---|---|---|---|
| minimumBid | minimum bid of the auction | 40.9 | 79 |
| Reserve | 1 if seller set a reserve price for the auction; 0 otherwise | 0.035 | 0.183 |
| sellerFeedback | Sellers' feedback score at time of listing | 44074.8 | 93126.7 |
| Duration | Duration of auctions in days | 4.12 | 2.6 |
| *Control variables: camera type, brand, condition, and product lines.* | | | |

## Results and Sample Size

We hypothesize according to the results of LBPR (hypotheses 1, 2, and 4 are directional, and hypothesis 3 is nondirectional[5]):

**Hypothesis 1**: Higher minimum bids lead to higher final prices.

**Hypothesis 2**: Auctions with reserve price will sell for higher prices.

**Hypothesis 3**: Duration affects price.

**Hypothesis 4:** The better the seller feedback, the higher the prices.

As discussed earlier, we investigate how the statistical significance of the parameters of interest changes as the sample size grows. The procedure, detailed below, is to re-estimate Equation [4] by adding a random sample of new observations to obtain the next sample for estimation. Our first

[5] LBPR hypothesized and found that longer duration was associated with higher prices. However, it seems reasonable that a short duration auction could increase bidding pressure and result in higher prices as well. Also, by not specifying the direction of the duration effect, we can illustrate a non-directional hypothesis with this example.

replication is to use Equation [4] and focus on the coefficient of the *duration* variable and the log transformation of *seller feedback score*. Figure 1 contains plots of the coefficient and of the p-value of duration as a function of the sample size; Figure 2 is the same plot for seller feedback score. We refer to these plots as coefficient/p-value/sample size plots, or *CPS plots* in the remainder of the paper. A general algorithm to create these graphs is:

For a sample of size *n*, and a CPS-plot based on *k* increasing sample sizes:

1. Choose the minimum sample size $n_0$ that is reasonable for fitting the model.
2. Randomly draw a sample of size $n_0$ from the large dataset
3. Fit the model of interest to this sample, and retain the estimated coefficients, their standard errors, and the p-values.
4. Increase the last sample size by adding *round(n/k)* more observations, drawn randomly from the remaining dataset.
5. Repeat steps (3)-(4) until the full original dataset is used.

Finally, create a line plot of the coefficients vs. the sample size (on the x-axis), and in another panel the p-value(s) vs. the sample size.

We implemented the above algorithm in Stata running 5000 iterations; sample Stata code is given in the appendix. Interested readers can make simple changes to the document, save it as a .do file, and run it from Stata to generate similar graphs for their own empirical studies.

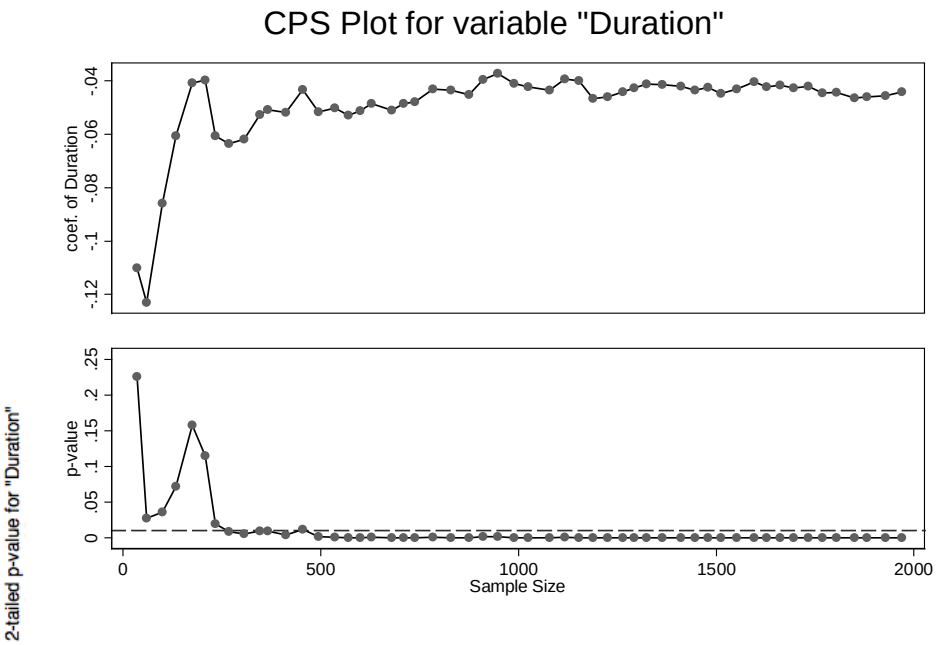


**Figure 1: CPS plot for "duration": coefficient and p-value vs. sample size. (Zoomed in to n < 2000 for illustration. Horizontal dotted line corresponds to p=0.01. )**

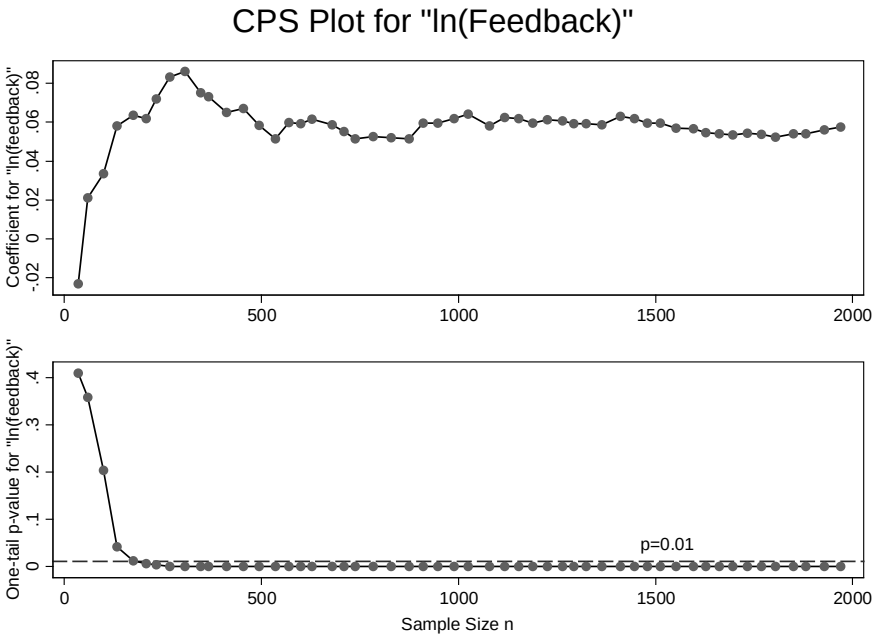


**Figure 2: CPS plot for ln(feedback): coefficient and p-value vs. sample size (Zoomed in to n < 2000 for illustration. Horizontal dashed line corresponds to p=0.01. )**

Figure 1 and Figure 2 show that once the sample size increases beyond some point, the p-value drops to near zero values and remains there. In this particular example, the p-value for the logarithm of seller feedback falls below 1% once the sample size is greater than 200; when the sample is larger than 500, the p-value is almost always less than 0.000001.

To further study the p-value distribution as a function of sample size, we use a Monte Carlo simulation to generate 400 samples for each sample size, for a set of increasing sample sizes. We then fit the same regression model, and compute the p-value for seller feedback. For example, we randomly sampled 100 data points from the full sample of camera auctions 400 times, ran the regression model on each of these subsamples, and plotted the resulting coefficients and p-values. Figure 3 shows the estimated distribution of coefficients and p-values as a function of sample size. The top and middle panels are a more general view of the CPS plot (compare this to the bottom panel in Figure 2). The median coefficient value is stable across the different sample sizes, and its variability decreases in a meaningful way; for samples below n=500 the distribution covers the value 0, yielding statistical insignificance at traditional significance levels. The plots show decreasing noise in the coefficient estimation reflecting the power of an increasing sample size. We see that not only do levels of p-values decrease rapidly with sample size, but so does the variability in the p-value distribution. In other words, with a large sample we expect to consistently see very small p-values. The bottom panel of Figure 3 displays the same p-value information on a logarithmic scale, better showing the minuscule magnitude of p-values at n>700. Note that although the variability in the bottom panel appears to increase with sample size, it actually decreases when considering the logarithmic scale.

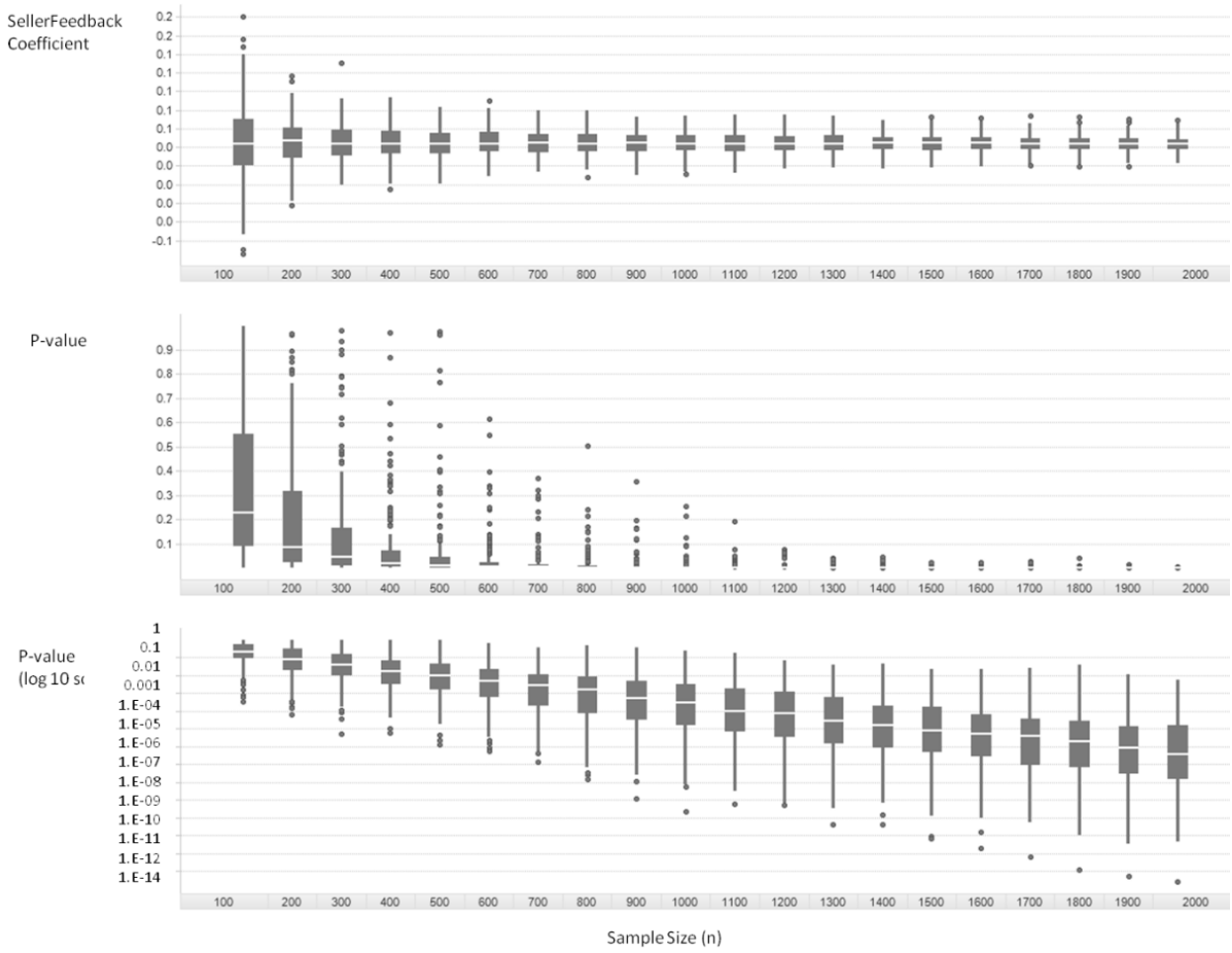


**Figure 3: Monte Carlo CPS Plot: Coefficient and p-value distributions as a function of sample size, using Monte Carlo simulation. The bottom panel displays the same p-value data on a logarithmic scale (note that apparently increasing variability is in fact decreasing due to the log scale). The white line within each box denotes the median.**

Figure 4 is a plot of the sample size at which each variable in Equation 4 becomes significant at the 1% level, which we call a *1% significance threshold plot*. Note that by an *n* of 300, all of the T values for testing the hypothesis that the coefficients are positive or different from 0 are in the rejection region. Beyond a sample size of 300, additional data drive down p-values and increase power. However, the p-value says nothing about the strength of the relationship between each independent variable and the dependent variable. Referring to Figure 2 again (and the top panel of Figure 3), the

coefficient for the relationship between feedback score and final price exhibits high variability until about 700 observations, but for samples over 700 up to the size of the full data set, the variability in the coefficient is low. It is this coefficient that measures the association between the independent and dependent variables, not a p-value.[6]

The impact of very large samples on p-values is a known problem in statistics and econometrics. Some econometricians have suggested that the critical value for determining significance, p*, be deflated to take it into account (Greene, 2002, p. 661; Leamer, 1978), however this suggestion does not work in practice. To test the feasibility of this idea, we used the "Seller feedback" variable to plot *p-value* versus *n*, and fit a curve to compute a deflated *p**. There are two problems with this approach. The first is that p* came out on the order of $10^{-6}$, a number that is virtually meaningless. Second, a practical problem is that most statistical packages do not report enough precision to calculate such extreme values.

We recognize that reviewers may require reports of statistical significance, as they are used to small-sample studies. If required to report significance measures, the researcher can employ Monte Carlo simulation to estimate significance levels using a series of smaller samples, as shown in the middle panel of Figure 3.

[6] The problem of misleading conclusions from low p-values is not confined to OLS. If a researcher chooses to use an instrumental variable (IV) to replace an endogenous independent variable in a regression, it is important that the IV is "sufficiently" correlated with the endogenous variable. Staiger and Stock (1997) show that large sample sizes could lead to an overestimation of the strength of instruments (Baum 2006).

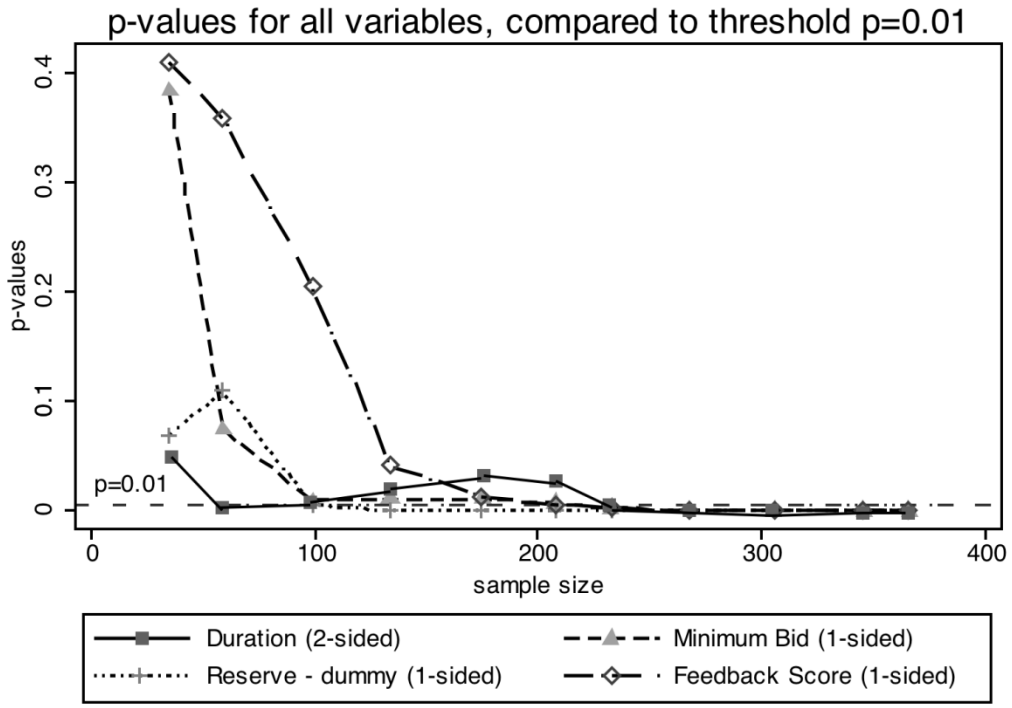


**Figure 4: p-values for four variables as compared with p* = 0.01 (horizontal line). (Zoomed in to n < 500 for illustration)**

Moving beyond statistical significance, a more important recommendation that we propose is to carefully examine the practical significance of variables. Table 3 illustrates the practical significance of the results; it shows how changes in the independent variables affect the dependent variable. Since 5% is the typical statistical significance threshold value used in empirical studies, we also present the 95% confidence interval using the entire data set, where the dependent variable is the log of price. In this example, the results are all statistically significant at the .0001 level, yet the influence of each independent variable on price is quite different. While a researcher would like to be able to make statements about the relative importance of each independent variable in influencing the dependent variable, there is much controversy over different approaches to solving this problem. Social scientists have commonly used beta weights or standardized coefficients that are independent of measurement

units. However, Goldberger and Manski (1995) argue that this technique is misleading. Other approaches have encountered criticism as well.

**Table 3: Practical Significance of Results**

| Variable | Coefficient (with entire dataset) | Standard Error | P-value | 95% Confidence Interval[7] | Coefficient Interpretation for the lower bound of the confidence interval |
|---|---|---|---|---|---|
| ln(minimum bid) | 0.1006 | 0.000825 | 0.000 | (0.0990, 0.1023) | 1% increase in the minimum bid is associated with an average 0.09% increase in final price, all else constant |
| Reserve | 0.7375 | 0.00675 | 0.000 | (0.7240, 0.7510) | Items with a reserve price sell for a price that is on average 106% (=100 ($e^{0.724}$ -1)%) higher, all else constant. |
| ln(feedback) | 0.0438 | 0.00065 | 0.000 | (0.0425, 0.0451) | 1% increase in the seller's feedback score is associated with an average of 0.04% higher price, all else constant |
| Duration | -0.0405 | 0.0007 | 0.000 | (-0.0419, -0.0391) | Each extra day for auction listing is associated with an average 4% decrease in price, all else constant |
| *Control variables*: conditions (used/new/refurbished, etc.); camera types; and brands | | | | | |

What we can say about the results in Table 3 is that if we hold three of the variables constant (along with the control variables), the coefficient of the fourth independent variable influences the

[7] It should be noted that while we use 95% for the confidence intervals, this is as subjective as the 5% cutoff for p-values. Also, for our directional hypotheses it would be more adequate to use one-sided confidence intervals (with a single lower or upper value); we present two-sided intervals due to their easier interpretation and popularity in IS.

dependent variable as indicated in the Table. The average selling price for the digital cameras in our dataset is $110.62, and the average of minimum bid is $89.52.

In Table 3, we calculate effect sizes using the lower bound of the confidence interval for each of the three variables with directional, positive hypotheses so we can say that a variable has *at least* that effect size with 95% confidence. For example, if we hold auction duration, seller feedback and reserve price constant, we are 95% certain (actually 97.5% because of directionality) that a 1% increase in the minimum bid is associated with *at least* an average 0.099% increase in the selling price. Similarly holding the other three independent variables of interest and control variables constant, we are 97.5% confident that a 1% increase in the seller's feedback score is associated with at least an average of a 0.04% higher price; and, each extra day for an auction listing, holding the other three independent variables constant, is associated with at least an average 4% decrease in price. Whether these magnitudes are practically significant in each context should be the focus of the researcher's attention.

One might argue that setting a reserve price has a large impact on the closing price of the auction. However, interpreting this result requires a better understanding of how data are generated and how they enter our analysis. The data in this study are for auctions that closed, that is, in which a buyer bought the product. Hence, the effect of reserve price on the final sales price is contingent on the fact that an auction closed, since no price information is available for unsold items. There are many auctions that do not close, often because of too high a reserve price. It would be misleading to argue that setting a very high reserve price will result in a much higher selling price, because the probability of an item being sold will be affected. To empirically estimate the overall effect of variables, we need different specifications of the model to adjust for such selection bias, such as Heckman models (Heckman 1979).

## How Are Results Misleading, and What is the Solution?

By using a large sample, IS researchers can demonstrate very impressive levels of statistical significance, which can mask the practical significance of their findings. Having a small p-value does not necessarily imply that a variable is important, and, when one variable has a smaller p-value than another, it is no guarantee that the first variable is more important than the second. As Table 4 shows, while the results for feedback, minimum bid and duration are highly statistically significant, these variables have a relatively small influence on the dependent variable. A researcher must ask: does a statistically significant variable have both economic and practical significance?

We suggest solutions to the problems that arise when testing hypotheses with very large samples, which should substantially improve the quality of information systems research. Our recommendations are intended to solve the p-value problem, provide readers with better evidence than the sign and direction of a regression coefficient, and encourage a sound presentation of the practical significance of findings.

- Do not rely on very small p-values and the direction of a coefficient as strong support for research propositions.
- Compute confidence intervals for the regression coefficients of interest. For a directional hypothesis, compute the more conservative bound of the confidence interval and present it as a conservative estimate of the minimum impact of the variable of interest with the certainty of the confidence interval.
- If required to present the traditional statistical significance of a variable, use Monte Carlo simulation to take repeated small samples and plot the coefficient and p-value distributions.
- Discuss the strength of the findings based on the context of the study.

We further expand on these recommendations below.

## 1) Direct Interpretation of Coefficients

Researchers should report the sensitivity of their dependent variable to changes in the independent variable in terms of *practical magnitude*, as illustrated in Table 3. Table 4 shows how to interpret effect sizes for a few of the most popular transformations in regression analysis. Focusing on effect sizes suggests that the reader should in fact ignore p-values for very large samples. Chatfield (1995) comments, "The question is not whether differences are 'significant' (they nearly always are in large samples), but whether they are interesting. Forget statistical significance, what is the practical significance of the results?"

**Table 4: Interpreting Effect Sizes for Common Regression Models (Vittinghoff et al., 2005)**

| **Functional Form** | **Effect Size Interpretation (where β is the coefficient)** |
|---|---|
| **Linear *f*** | |
| *y = f(x)* | A unit change in x is associated with an average change of β units in y |
| *ln(y)=f(x)* | For a unit increase in x, y increases on average by the percentage100($e^{\beta}$-1) (≅100 β when \| β\|<0.1) |
| *y=f(ln(x))* | For a 1% increase in x, y increases on average by ln(1.01)* β (≅β/100) |
| *ln(y)=f(ln(x))* | For a 1% increase in x, y increases on average by the percentage100($e^{\beta *\ln(1.01)}$-1) (≅β when \| β\|<0.1) |
| **Logistic *f*** | |
| Numerical x | A unit change in x is associated with an average change in the odds of Y=1 by a factor of β. |
| Binary x | The odds of Y=1 at x=1 are higher than at x=0 by a factor of β |

Marginal analysis extends our discussion of effect sizes beyond linear and exponential / multiplicative models. In our OLS example, the marginal effect is the same for any X value. However, when dealing with models like the probit, one has to specify whether an effect size is being calculated at

the mean of X or some other value such as the median. For example, assume one did a probit analysis and wanted to interpret the coefficient for a variable $X_1$. The researcher would hold all of the other Xs at a certain value such as their median, and then measure the change in Y as a function of increasing $X_1$ by a unit.

In addition, for nonlinear models, marginal analysis is a more robust way – and sometimes the only way — to interpret the effect size, compared to looking at the p-value or even magnitude of the coefficient. As an example, if we have $X_1$ and $X_1^2$ as the explanatory variables for *Y*, it is incorrect to directly interpret the marginal effect of $X_1$ solely based on its coefficients, because we cannot hold $X_1^2$ constant and at the same time increase $X_1$ by one unit.

## 2) Confidence Intervals for Reporting Purposes

We recommend that IS researchers working with large samples report effect sizes using confidence intervals, an approach often used in empirical economics research (Goolsbee and Guryan, 2006; Disdier and Head, 2008; Cannon and Cipriani, 2006). Whereas the p-value only describes the probability that the null hypothesis (e.g., coefficient equals 0) can be rejected given a true effect, the confidence interval (CI) estimate gives us a range for the actual magnitude of the parameter of interest. As the sample size increases, a typical CI will become narrower around the point estimate, such that at some point it effectively coincides with the point estimate. In other words, while the information that p-values convey does not scale up to large samples, the information contained in confidence intervals does. This property means that even if the researcher is unsure whether the sample is too large for using p-values, relying on a CI is always safe. See Figure 5 for an example of confidence intervals from the digital camera auctions analysis. For each sample size the confidence interval was computed as the coefficient +/- 1.96 times the standard error. Furthermore, we advocate using the most conservative bound of the confidence interval and reporting that the researchers are, for example, 95% confident

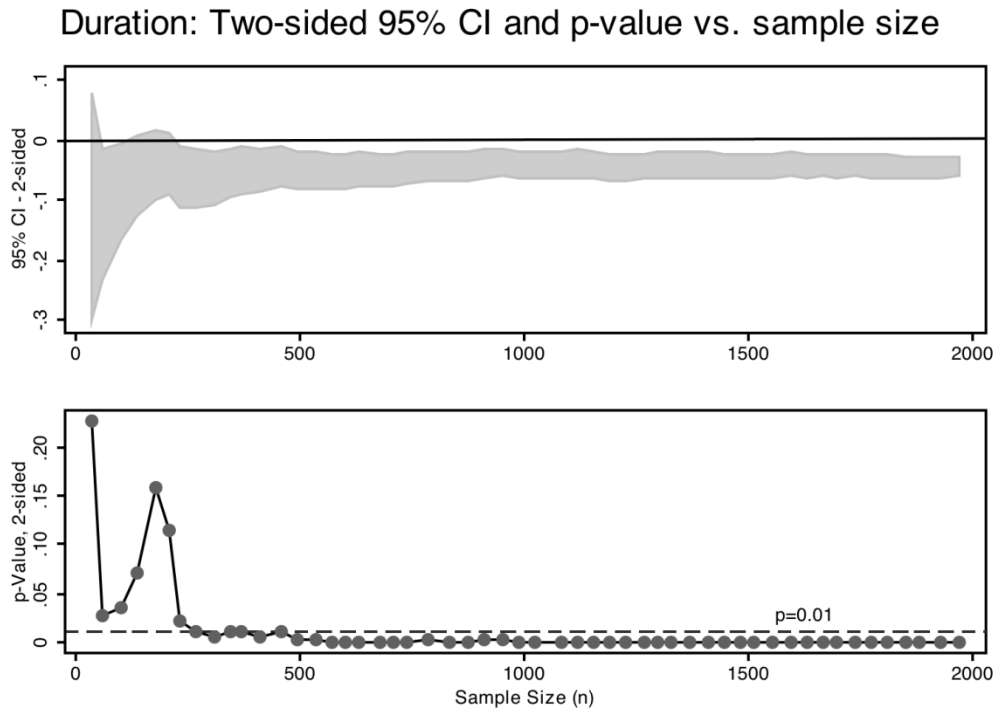


**Figure 5: Two-sided 95% confidence interval (top) and p-value (bottom) for duration vs. sample size. (Zoomed in to n < 2000 for illustration. Horizontal dotted line in the bottom panel corresponds to threshold value p=0.01. )**

that the independent variable has the calculated impact on the dependent variable.

There are four additional major benefits to reporting confidence intervals over p-values and coefficient signs. First, confidence intervals solve the problem that motivated this paper: the tendency of large-sample IS research to rely on low p-values and the direction of a regression coefficient to support the researcher's propositions. Second, when researchers report the confidence intervals for a particular variable across different studies, it becomes much easier to conduct meta-analysis, synthesize prior studies, and help advance scientific knowledge of the relevant IS field. This is particularly true when the CI is for elasticities (that is, the percentage change in Y for each 1% change in X). An example can be found in de Leeuw (1971) from the labor economics literature. The third benefit is that, when researchers have doubts about their data or methodology, they can use a CI to interpret the results in a more conservative way, thereby lending more credibility to their findings. Last but not least, empirical IS

researchers tend to conduct multiple robustness checks for their models by comparing multiple model specifications. With CIs, we can go beyond the argument that "results are qualitatively similar", and quantitatively compare the range of estimates. Each specification will provide a CI, and we can use the degree to which these CIs overlap as a criterion for quantitative consistency. If the robustness test yields a wildly different CI, it will be certainly worth some further exploration. Examples of the use of a CI or one of its bounds can be found in many papers in many empirical fields (Iglesias and Riboud 1988; Goolsbee 2000; Black et al. 2002; Vissing-Jørgensen 2002; Goolsbee and Guryan 2006).

Hubbard and Armstrong (2006) offer a number of additional arguments in favor of confidence intervals over other forms of reporting. A confidence interval provides a range of estimates that the research suggests is likely for the population, and it provides a measure of the reliability or precision of the estimate. The confidence interval makes it easier to assess substantive rather than statistical significance. If a reader is fond of testing hypotheses, as an example, a 90% confidence interval that fails to include the null value (usually 0 for a coefficient in a regression equation) is equivalent to rejecting the null hypothesis at the .05 level (for the non-directional case).

## 3) Conservative Claims of Statistical Significance

Many readers and reviewers are used to reports of statistical significance in studies with small sample sizes and may be uncomfortable with arguments that p-values are not meaningful in very large samples. If a researcher is required to respond to this criticism, we recommend following the Monte Carlo sampling approach described earlier. By estimating a model 400 times with small sub-samples randomly drawn from the large sample, one can plot the resulting coefficient and p-value distributions as a function of sample size as shown in Figure 3.

## 4) Strength of Findings

Readers and reviewers of earlier drafts of this paper challenged us to suggest effect size strength measures similar to those proposed by Cohen (1988) even though his purpose is different from ours. Cohen defines a statistic $f^2$ for regression analysis as $R^2/(1-R^2)$ and suggests a value of .02 as small, .15 as medium and .35 as strong. This variance-based measure may be useful in the social sciences when there are no natural metrics for a variable, for example, when the researcher uses a Likert scale of 1 to 5 or 1 to 7. The $f^2$ statistic is dimensionless as is Cohen's effect size strength scale. In most of the large-sample IS studies we reviewed, the variables in the analysis have easily interpretable metrics like dollars or numbers of visits to a web site as opposed to a Likert scale (Olejnik and Algina , 2000; Kortlik and Williams, 2003). Cohen's purpose is to determine the sample size needed to have adequate power in a study given the researcher's assumptions about the strength of the relationship. Our purpose in discussing effect size is to see whether or not a specific independent variable has any practical impact on the dependent variable.

We have been reluctant to offer strength measures for findings because strength depends on the context of the study. As an example, a sensitivity analysis of the results of the digital camera auctions shows that approximately a 1% increase in the minimum bid is associated with a 0.1% increase in final sale price all else held constant. The effect of this finding given the $110 average price of cameras in the auction is 11 cents, which is of limited practical significance. However, if the dependent variable is wireless spectrum for cellular communications with an average selling price of $2 billion, the same coefficient with a 1% change in minimum bid produces an effect size of $2 million, which could be of practical significance to a bidder.

With this caveat in mind we propose the following guidelines when presenting large sample IS research with an easily interpreted measurement for the dependent variable like dollars. First, the

researcher should conduct a sensitivity analysis similar to the one above for digital cameras. In the case of the log transformation in our example, the percentage changes are obvious. For other functional forms, compute sensitivity for a 1% change in the independent variable. For example, in an OLS estimate with no transformations, take 1% of the mean value of the independent variable times its coefficient and divide the results by the mean of the dependent variable to obtain the percentage change in y associated with a 1% change in x.

The objective of conducting this sensitivity analysis is to assess the practical significance of an independent variable keeping in mind the earlier comments about study context. Distinguishing between small and large effects is strongly context dependent. For a small effect, the question we asked is at what point would we be willing to recommend taking action to a decision maker based on the results of a study? It is unlikely that a manager would be interested in a variable that has a 1% or less impact on a dependent variable even if she believed there was a strong causal relationship. At the other extreme, given that large sample research usually features a large number of independent variables of interest as well as control variables, when would a researcher feel that one of these variables has a strong effect on the dependent variable? No matter the numbers, researchers should make a case for the strength of the effect size and readers should evaluate those arguments given the context of each study.

## 5) Improved Model Validation in Large Sample Studies: Using Monte Carlo

Valid interpretations of individual effects from regression models require that the underlying model assumptions are met. Given a large sample, researchers can take advantage of the wealth of data to easily test model assumptions regarding coefficient distributions, and even obtain valid coefficient interpretations if the assumptions are violated, via an approach similar to the Monte Carlo resampling that we presented earlier. According to this approach, the researcher computes the empirical

distribution of coefficients (or any other statistic of interest) using a set of randomly divided, nonoverlaping subsamples of the data for each variable of interest. For example, with a sample size of 100,000, a researcher could estimate the model on subsets of 1,000 data points (assuming that a sample of size 1,000 is sufficient for model estimation) and compare the resulting empirical coefficient distributions to those expected under model assumptions (e.g., normally distributed).

To determine the size of the subsamples to use, we recommend first estimating a model on the full dataset, and then generating the CPS plot to see at what subsample size the coefficients exhibit stability and p-values are close to zero. Then divide the full sample into subsamples of approximately this size and re-estimate the model for each subsample. Finally, plot the distribution (e.g., using histograms, boxplots, or probability plots) of the computed coefficients from the multiple subsamples, for each coefficient related to a hypothesis, thereby providing the empirical distribution of the estimates of interest. The plot can detect whether the relevant model assumptions appear to hold, such that confidence intervals based on the estimated model from the full sample may be properly interpreted. Otherwise, the empirical distributions of the coefficients can be used for hypothesis testing by finding, say, the corresponding $5^{th}$ and $95^{th}$ percentiles.

As an example, for our dataset of camera sales, we divided the full sample into 999 subsamples and estimated the model 999 times, producing the plots in Figure 6 and the results in Table 5. The standard deviations of the coefficients from the subsamples compare very closely with the standard errors estimated from the model for the full sample. The histograms of the coefficients, based on the 999 samples, also imply normally distributed coefficients, as assumed by OLS. Thus, we can use the standard errors in the full model to construct confidence intervals for the coefficients. If, however, the subsample results had deviated from the full model results, or indicated a violation of assumptions, then we could use the subsample results for hypothesis testing and constructing empirical confidence

intervals.

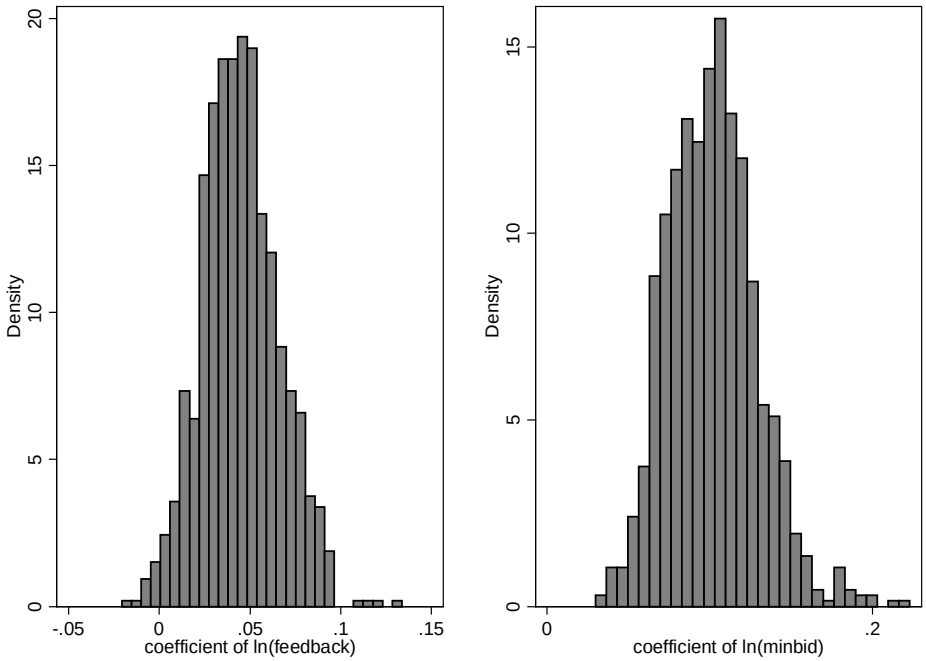


**Figure 6: Plots of log of feedback and log of initial bid from estimating the model on 999 subsamples**

**Table 5: Comparison of Confidence Intervals from the Full Sample**

| **Variable** | **95% CI from overall sample** | **95% from the subsample estimate** |
|---|---|---|
| ln(minbid) | (0.0990, 0.1023) | (0.0998, 0.1032) |
| Reserve | (0.7240, 0.7510) | (0.7287, 0.7591) |
| ln(feedback) | (0.0425, 0.0451) | (0.0434, 0.0460) |
| Duration | (-0.0419, -0.0391) | (-0.0424, -0.0393) |

## Conclusions

In this paper we argue that if a sample is "too large" then p-values are likely to be useless and potentially misleading; the researcher should concentrate on the meaning of the coefficients. When the sample size is large and a p-value indicates statistical significance, we recommend ignoring the p-value and instead reporting the effect size for the variable using the lower and/or upper bound of the variable's confidence interval (depending on the hypothesis direction). The important information from the study is how much influence the independent variable has on the dependent variable. In addition, reporting confidence intervals makes it possible to compare findings across different studies.

The availability of very large samples is changing the nature of IS research, especially on consumers and markets. Our concern is that reliance on statistical significance in testing models on large samples can mislead readers and lead to erroneous conclusions that can have substantial practical impact. The strength of the findings should rest not on a low p-value, but instead on the practical significance of the estimated parameters. We strongly recommend that IS researchers reduce their emphasis on p-values with large samples and that they present confidence intervals and the effect size of the *actual* coefficient magnitudes. We believe that these steps will strengthen the findings and credibility of information systems research conducted with large samples.

## Acknowledgements

The authors thank Professor Foster Provost from NYU for planting the seeds for this paper and Professor Wolfgang Jank from UMD for sharing the digital camera eBay data.

## Appendix A: Why the P-Value Approaches 0 for Large Samples

Traditionally, empirical research papers in IS explicitly discuss and elaborate on what a statistician would call the alternative hypothesis, for example, that females use smart phones for texting more than males or that a higher starting bid in an online auction is associated with a higher final price for the goods being auctioned, implicitly implying that the null hypothesis is the "opposite" scenario. Underlying all statistical testing is the null hypothesis which always includes the case of "no effect," for example, that there are no differences between groups or that there is no association among variables.

The null hypothesis either contains only the non-directional "no effect" scenario or it contains both the "no effect" scenario and the "opposite" directional scenario. The researcher hypothesizing that a coefficient in a regression equation is positive is trying to reject the null hypothesis that the coefficient is 0 or negative, i.e., rejecting the no effect and negative coefficient scenarios. When a researcher reports that the coefficient of the regression equation is positive and statistically significant at the .05 level, there is only a 5% chance that she would have observed this result or one more extreme (i.e., a larger positive coefficient) if in fact the coefficient is 0 or negative.

Consider a researcher who conducts an online survey of college students to test the alternative hypothesis that female students use their smart phones for texting more than male students. The implied null hypothesis is that either there is no gender effect or that male students use their smart phones for texting more than female students. The researcher's survey displays a line on the respondent's computer anchored by 0% and 100% on either end, and asks the respondent to click at the percentage of their smart phone use that is for texting. If male and females actually text the same amount, the first problem is accurately measuring the responses from the continuous line where the respondents click on their responses. As Tukey (1991) put it: "Are the effects of A *(females)* and B *(males)* different? They are always different---for some decimal place." Cohen (1990) says "A little thought reveals a fact widely understood among statisticians: The null hypothesis, taken literally (and that's the only way you can take it in formal hypothesis testing), is always false in the real world….If it is false, even to a tiny degree, it must be the case that a large enough sample will produce a significant result and lead to its rejection. So if the null hypothesis is always false, what's the big deal about rejecting it?" We are not suggesting that IS researchers abandon hypothesis testing; these observations on the null hypothesis are intended to move our focus from relying solely on statistical significance to consideration of practical significance and effect size.

Large samples are advantageous because of the statistical power that they provide. Yet researchers should also realize that a by-product of increasing the sample size is that the p-value itself will easily go to zero. The p-value for testing a non-directional hypothesis regarding a linear regression coefficient is calculated by:

[1] $p\text{-value} = 2 * (1 - \Phi(df, |T|))$

where $\Phi$ is the cumulative student's t-distribution, *df* is the residual degrees of freedom, and $|T|$ is the absolute value of the observed t-statistic[8], given by $|T| = \frac{\hat{\beta} - 0}{\hat{\sigma}_{\hat{\beta}}}$.

This T statistic is an increasing function of the sample size *n,* because the standard error in the denominator decreases in *n*. In the case of a single independent variable, it is straightforward to see the effect of the sample size on the standard error:

$$\hat{\sigma}_{\hat{\beta}} = \frac{\sqrt{MSE}}{s_x\sqrt{n-1}}$$

where MSE is the estimate of the error variance and $s_x$ is the standard deviation of the independent variable ( $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$ ). Hence, in the single independent variable case we can write |T|= $\left|\frac{\hat{\beta} - 0}{\frac{\sqrt{MSE}}{s_x\sqrt{n-1}}}\right| = \sqrt{n-1}\frac{\left|\hat{\beta}\right| \times s_x}{\sqrt{MSE}}$

What happens to p-value as *n*, the sample size, increases? Consider the null hypothesis $H_0$: $\beta$=0. Unless the null hypothesis is *exactly* true (to an infinite number of decimals), the p-value will go to 0 as *n* becomes infinitely large, because the value of |T| will approach infinity, and therefore the cumulative t distribution until |T| (which becomes effectively a standard normal distribution) approaches 1. Equation [2] shows the limit of the p-value for the cumulative t distribution used to determine the statistical significance of a regression coefficient in the case of a single independent variable:

[2] $\lim_{n\to\infty} p\text{-value} = \lim_{n\to\infty} 2*(1-\Phi(df,|T|)) = 2*(1-\lim_{n\to\infty}\Phi(df,|T|)) = 0$

$$\lim_{n\to\infty} \text{p-value} = \lim_{n\to\infty} 2\times\left(1-\Phi(df,|T|)\right) = 2\times\left(1-\lim_{n\to\infty}\Phi(df,|T|)\right) = 2\times\left(1-\Phi\left(df,\lim_{n\to\infty}\sqrt{n-1}\frac{\left|\hat{\beta}\right|\times s_x}{\sqrt{MSE}}\right)\right) =$$

$$= \begin{cases} 2\times\left(1-\Phi(df,0)\right) = 1, & \text{if } \beta = 0 \\ 2\times\left(1-\Phi(df,\infty)\right) = 0 & \text{if } \beta \neq 0 \end{cases}$$

[8] Source: Buis, M. L. (2007). Stata tip 53: Where did my p-values go? *Stata Journal, 7*(4), 584-586.

Note that unless β is exactly equal to 0 with an infinite number of decimals (in which case the p-value will approach 1), the p-value will approach 0. A similar mathematical relationship exists between the test statistic and the sample size in *all statistical tests*, including regression models with multiple independent variables, t-tests, ANOVA, etc. It is easy to understand this if we think of the sample size approaching the entire population. If we know the exact value of β (or another parameter of interest) in the population, we also know whether it is exactly equal to 0 (or a different value of interest) or not with no uncertainty.

Many IS papers utilizing regression models, test that a coefficient is either positive or negative (directional hypothesis) and evaluate statistical significance with a one-sided test. The illustration above is for two-sided tests and can be modified for a one-sided test by eliminating the 2's in Equations 1 and 2, replacing |T| with T, and for a negative coefficient hypothesis (H1: $\beta<0$), replacing 1-$\Phi$ with $\Phi$. At the limit, these changes have no effect on the p-value approaching zero or 1 in large samples.

This artificial deflation of the p-value as sample size increases is well known in statistics (e.g. Chatfield 1995). When one has 500,000 observations, the p-values associated with the coefficients from modeling this data set are almost always going to be 0, so that a statistical test is close to useless at best, and misleading at worst. Econometricians have also long realized this issue and suggest that the threshold p-value should be adjusted downwards as the sample size grows (Leamer 1978; Greene 2003), however to our knowledge there have been no proposed rules-of-thumb in terms of how such adjustments should be made.

*This fascination with p-values comes because researchers too often confuse p-value with effect size.* In a conventional test of a hypothesis, a researcher establishes the criterion for accepting or rejecting the null hypothesis before collecting a sample. If she chooses the 5% level, it means that if her test statistic is in the rejection region, there is only a 5% chance she would obtain this test statistic if the null hypothesis is true. If, instead, she chose the 1% level and the test statistic is in the rejection region, then there is only a 1% chance she would get this result if the null hypothesis is true. *The p-value indicates the probability that one would observe the test statistic (or a more extreme value) given the null hypothesis is true.* The *p*-value says nothing about the *strength* of the effect under investigation. A p-value<.001 does not imply a stronger relationship between variables than a p-value<.01.

As an example, Thompson (1989) presents a table of results with fixed effect sizes showing increasing levels of statistical significance as the sample size increases. The level of statistical significance increases, but the strength of the relationship in the table remains constant. The result becomes statistically significant somewhere between 13 and 23 observations in the sample, but the effect size is fixed.

Researchers in many fields seem to regard a test statistic that allows them to reject the null hypothesis at the 5% level as magical proof of the relationship they believe exists between independent and dependent variables. A focus on a particular level of significance has led to suggestions that we have become so obsessed with 5% that we have forgotten to look at the practical significance of our findings (Ziliak and McCloskey, 2007; Sawyer and Peter, 1983; Carver, 1978).

Another related issue is the meaning of statistical *insignificance* in a large-sample study. If an effect remains statistically insignificant even with a very large amount of data, it provides evidence in favor of the null hypothesis. This is especially likely in directional hypotheses, where the null hypothesis includes both the "no effect" and "opposite effect" scenarios. Because researchers typically do not look at the magnitude of statistical *insignificance*, there is no danger of incorrect inference (as in the case of inference from low p-values). However, a statistically insignificant effect should be examined and

interpreted carefully in terms of its magnitude, to see whether it reflects a “no effect” or “opposite effect” scenario.

## Appendix B: Sample Stata code to generate CPS plot

```
/* Beginning of file */

/* This file contains the Stata code for estimation and generation of CPS plots.  First, please
specify the location for the output (log file and datasets with p-values etc. and sample size) as
well as the dataset you are analyzing. Please define your dependent variables on the lines
“global m” and “global y” first, then change the model you estimate (the line marked with “***
define your model here ***”) and be sure that you dataset does not have a variable named
“sampling”.  The following sample file collects the p/t/se/n/b of one explanatory variable of
interest, as sample size increases. */

Clear
version 9
capture log close
* set mem 600m /* change this to a size sufficient for your own data */
set matsize 5001
set more off

* specify the location here (where the log file and the final dataset will be saved)
cd “”

* Change your dataset location
use "C:\ebaydata.dta", clear

* Specify the *one* variable that you’d like to plot the CPS plot
global m "lnfeedback"
log using cps_$m.smcl, replace

* defining entire explanatory variable list EXCLUDING the one defined above
global x "reserve lnminbid duration condition2-condition6  cameratype2-cameratype7 brand2-
brand33"

* dependent variable
global y "lnprice"

* this is the number of points to plot on the CPS plot (here it is ln(feedback))
local z = 5000



* defining matrices to hold these values: b, p, t, se, n
matrix bvals = J(`z', 1, 0)
matrix pvals = J(`z', 1, 0)
matrix sevals = J(`z', 1, 0)
matrix tvals = J(`z', 1, 0)
matrix nvals = J(`z', 1, 0)

set seed 12345
capture gen sampling = uniform()
local j = 0

forvalues i=0.0002(0.0002)1  {
       local j = `j' + 1
       disp `j'
```

```
        ***  Define your model here ***
        qui reg $y $m $x if sampling <= `i', robust
        matrix betas = e(b)
        matrix bvals[`j',1] = betas[1,1]
        matrix nvals[`j',1] = e(N)
        matrix vc = e(V)
        matrix sevals[`j',1] = sqrt(vc[1,1])
        qui testparm $m
        matrix pvals[`j',1] = r(p)
        * Divide the p-value by 2 for one-sided tests. Remove "/2" for two-sided tests.
        * matrix pvals[`j',1] = r(p)/2
        matrix tvals[`j',1] = betas[1,1] / (sqrt(vc[1,1]))
}

drop _all

* now create a new dataset to include all the results from the iterations
svmat nvals
svmat bvals
svmat pvals
svmat sevals
svmat tvals

local parlist "n b p se t"
foreach z of local parlist {
        rename `z'vals1 `z'_$m
}
save CPS_$m.dta, replace

* now generate a dual-panel CPS plot

scatter b_$m n if n<= 2000, connect("l") yline(0.01) scheme(s1mono) saving(xxx.gph)
scatter p_$m n if n<= 2000, connect("l") scheme(s1mono) saving(yyy.gph)
graph combine xxx.gph yyy.gph, col(1) saving(cps_$m.gph)

erase xxx.gph
erase yyy.gph

log close
set more on

/* end of file */
```

# Appendix C: Sample Stata codes to generate Monte Carlo plot of p-values using modest subsamples

```
/* Beginning of file */

clear
version 11
capture log close
set mem 600m
set more off

* specify the location here (where the log file and the final dataset will be saved)
cd "C:\sample\mc\"
* Change your dataset location
```

```
use "C:\sample\eBay.dta", clear

* open the large sample data file

use "C:\ebay.dta"

capture snapshot erase _all

* Specify the *one* variable that you'd like to plot the CPS plot
global m "lnfeedback"

* defining entire explanatory variable list EXCLUDING the one defined above
global x "lnminbid reserve duration condition2-condition6 cameratype2-cameratype7 brand2-brand33"

* dependent variable
global y "lnprice"

* Keep only valid data points before sampling (i.e. removing rows with missing values)
qui reg $y $m $x
gen byte insample = (e(sample))
keep if insample == 1

* this is the number of points to plot on the CPS plot (here it is ln(feedback))
local z = 200
* defining matrices to hold these values: b, p, n
matrix pvals = J(`z', 1, 0)
matrix nvals = J(`z', 1, 0)
matrix bvals = J(`z', 1, 0)

snapshot save

* Define the sample size for each step

forvalues ss=100(100)2000{

        snapshot restore 1
        local j = 0
        disp "******************** subsample size is now: " + `ss'

        forvalues i=1(1)200{
                set seed `i'
                sample `ss', count
                local j = `j' + 1
                disp `j'
                        ***   Define your model here ***
                        qui reg $y $m $x , robust
                        matrix nvals[`j',1] = e(N)
                        matrix betas = e(b)
                        matrix bvals[`j',1] = betas[1,1]
                        qui testparm $m
                        matrix pvals[`j',1] = r(p)

                snapshot restore 1
        }

        drop _all

        * Now create a new dataset to include all the results from the iterations
```

```
		svmat nvals
		svmat pvals
		svmat bvals

		summ pvals1
		save ss_`ss'.dta, replace

}

** end of double-loops; to combine all variables

clear
use ss_100.dta

forvalues i=200(100)2000{
		append using ss_`i'.dta
		}

save ss_all.dta, replace

* Graph and save in the local folder

graph box p, over(n) nooutsides saving(p_n.gph)
graph box b, over(n) nooutsides saving(b_n.gph)

* You can now use “snapshot restore 1” to return to your original dataset.

/* End of codes */
```

# References


Baum, C. *An Introduction to Modern Econometrics using STATA,* Stata Corp, 2006.

Black, D., Daniel, K., and Sanders, S. "The Impact of Economic Conditions on Participation in Disability Programs: Evidence from the Coal Boom and Bust," *American Economic Review* (92:1) 2002, pp 27-50.

Cannon, E, and G. P. Cipriani, “Euro-Illusion: A Natural Experiment, *Journal of Money, Credit & Banking,* Vol. 38, No. 5 (August 2006), pp. 1391-1403.

Carver, R.P. "The Case Against Statistical Significance Testing," *Harvard Educational Review* (48:3) 1978, pp 378-399.

Chatfield, C. *Problem Solving: a Statistician's Guide* Chapman & Hall/CRC, 1995.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* Lawrence Erlbaum, 1988.

Cohen, J. "Things I have learned (so far)," *American Psychologist* (45:12) 1990, pp 1304-1312.

de Leeuw, F. "The Demand For Housing: A Review Of Cross-Section Evidence," *Review of Economics & Statistics* (53:1) 1971, p 1.

Disdier, A. and K. Head. "The Puzzling Persistence of the Distance Effect on Bilateral Trade, *Review of Economics & Statistics*, Vol. 90, No. 1 (Feb. 2008), pp. 37-48.

Forman, C., Ghose, A., and Wiesenfeld, B. "Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets," *Information Systems Research* (19:3), September 1, 2008 2008, pp 291-313.

Ghose, A., Smith, M., and Telang, R. "Internet Exchanges for Used Books: An Empirical Analysis of Product Cannibalization and Welfare Impact," *Information Systems Research* (17:1) 2006, p 3.

Goldberger, A.S., and Manski, C.F. "The Bell Curve by Herrnstein and Murray," *Journal of Economic Literature*), Jan 1 2008.

Goldfarb, A., and Lu, Q. "Household-Specific Regressions using Clickstream Data," *Statistical science* (21:2) 2006, pp 247-255.

Goolsbee, A. "What Happens When You Tax the Rich? Evidence from Executive Compensation," *Journal of Political Economy* (108:2) 2000, p 352.

Goolsbee, A., and Guryan, J. "The Impact of Internet Subsidies in Public Schools," *Review of Economics & Statistics* (88:2) 2006, pp 336-347.

Greene, W. *Econometric analysis* prentice Hall Upper Saddle River, NJ, 2002.

Heckman, J. "Sample Selection Bias as a Specification Error," *Econometrica* (47:1) 1979, pp 153-161.

Hubbard, R., and Armstrong, J. "Why We Don't Really Know What Statistical Significance Means: a Major Educational Failure," *Journal of Marketing Education* (28) 2006, pp 114-120.

Iglesias, F.H., and Riboud, M. "Intergenerational Effects on Fertility Behavior and Earnings Mobility in Spain," *Review of Economics & Statistics* (70:2) 1988, p 253.

Kotrlik, J., and Williams, H. "The Incorporation of Effect Size in Information Technology, Learning, Information Technology, Learning, and Performance Research and Performance Research," *Information Technology, Learning, and Performance Journal* (21:1) 2003, p 1.

Leamer, E. *Specification Searches: Ad Hoc Inference with Nonexperimental Data* John Wiley & Sons Inc, 1978.

Lucking-Reiley, D., Bryan, D., Prasad, N., and Reeves, D. "Pennies from eBay: The Determinants of Price in Online Auctions," *The Journal of Industrial Economics* (55:2) 2007, pp 223-233.

Manning, W.G., Duan, N., and Rogers, W.H. "Monte Carlo evidence on the choice between sample selection and two-part models," *Journal of Econometrics*), Jan 1 1987.

Olejnik, S., and Algina, J. "Measures of Effect Size for Comparative Studies: Applications, Interpretations, and Limitations," *Contemporary Educational Psychology* (25:3) 2000, pp 241-286.

Overby, E., and Jap, S. "Electronic and Physical Market Channels: A Multiyear Investigation in a Market for Products of Uncertain Quality," *Management Science*) 2009.

Pavlou, P., and Dimoka, A. "The Nature and Role of Feedback Text Comments in Online Marketplaces: Implications for Trust Building, Price Premiums, and Seller Differentiation," *Information Systems Research* (17:4) 2006, pp 392-414.

Sawyer, A., and Peter, J. "The Significance of Statistical Significance Tests in Marketing Research," *Journal of Marketing Research* (20:2) 1983, pp 122-133.

Staiger, D., and Stock, J. "Instrumental Variables Regression with Weak Instruments," *Econometrica: Journal of the Econometric Society* (65:3) 1997, pp 557-586.

Thompson, B. "Statistical Significance, Result Importance, and Result Generalizability: Three Noteworthy But Somewhat Different Issues," *Measurement and evaluation in Counseling and Development* (22:1) 1989, pp 2-6.

Tukey, J. "The Philosophy of Multiple Comparisons," *Statistical science* (6:1) 1991, pp 100-116.

Vissing-Jørgensen, A. "Limited Asset Market Participation and the Elasticity of Intertemporal Substitution," *Journal of Political Economy* (110:4) 2002, pp 825-853.

Vittinghoff, E., and Glidden, D. *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models* Springer Verlag, 2005.

Wu, J., Cook Jr, V., and Strong, E. "A Two-Stage Model of the Promotional Performance of Pure Online Firms," *Information Systems Research* (16:4) 2005, p 334.

Ziliak, S., and McCloskey, D. *The Cult of Statistical Significance* University of Michigan Press, 2008.