STA 371H: Final Exam

May 13, 2015

Statistics and Modeling

Instructions:

1. Do not turn this page and begin the exam until instructed to do so.
2. Write in ink, either in a blue book or on separate sheets of paper. If you are not using a blue book, write your name and UT EID on each page, and staple all pages together. I will not grade anything written on the exam sheet itself, so you may use this as scratch paper.
3. Turn in this exam paper along with your written exam.
4. This is a closed-book exam; you are allowed pens and paper on your desk, and nothing more.
5. Switch off all cell phones, mobile communication devices, iPods, and so forth. Do not merely turn them to silent or vibrate mode.
6. The time limit for this exam is 180 minutes. Budget your time wisely.
7. The exam has three parts with point values labeled. These add up to 140 points.
8. On the following page I have provided several formulas we have used this semester.

Good luck!

Here are some formulas that you may find helpful.

Addition rule: The probability that either A or B will happen is

$$P(A \text{ or } B) = P(A) + P(B) - P(A,B),$$

where $P(A,B)$ is the probability that both A and B happen at once.

Multiplication rule: The joint probability that A and B will both happen is

$$P(A,B) = P(A) \cdot P(B \mid A),$$

where $P(B \mid A)$ is the conditional probability that B will happen, given that A happens.

Bayes' rule: a rule for updating prior probabilities into posterior probabilities, given new data.

$$P(A \mid B) = \frac{P(A) \cdot P(B \mid A)}{P(B)}.$$

Moments. If a random variable X has N possible outcomes $\{x_1, \ldots, x_N\}$ having corresponding probabilities $\{p_1, \ldots, p_N\}$, then the expected value is

$$E(X) = \sum_{i=1}^{N} p_i x_i.$$

The expected value of a function of X, denoted $f(X)$, is

$$E[f(X)] = \sum_{i=1}^{N} p_i f(x_i).$$

The variance of X is

$$\text{var}(X) = E\left(\{X - E(X)\}^2\right).$$

And if $P(X,Y)$ is a joint distribution of two random variables X and Y, then the covariance of X and Y is

$$\text{cov}(X,Y) = E\Big\{[X - E(X)][Y - E(Y)]\Big\}.$$

Linear combinations. Suppose that W is a linear combination

$$W = aX + bY + c$$

of two random variables X and Y, for some constants a, b, and c. The moments of the linear combination W can be described in terms of the moments of the joint distribution $P(X,Y)$.

$$\begin{aligned} E(W) &= aE(X) + bE(Y) + c \\ \text{var}(W) &= a^2 \text{ var}(X) + b^2 \text{ var}(Y) + 2ab \text{ cov}(X,Y). \end{aligned}$$

Part 1: Short answer (40 points)

A: Briefly describe what someone might be trying to accomplish by running a logistic regression. Why might one prefer logistic regression to the so-called "linear probability model"?

B: Suppose that X is a lottery that leaves you with either \$900 or \$1600, with equal probabilities. **True or false: If your utility function for wealth is $u(w) = \log w$, then your expected utility is $E(u\{X\}) = \log(0.5 \cdot 900 + 0.5 \cdot 1600) = \log(1250)$.** Either way, explain why.

C: Suppose you are trying to model the performance of a portfolio consisting of two assets. Let X and Y be random variables representing the returns of these two assets. You decide to model X and Y with a joint probability distribution, and to specify the following moments of the joint distribution to the best of your ability: the expected values and variances of both X and Y, along with the covariance of X and Y.

Your portfolio W is a linear combination of X and Y: that is, $W = aX + bY$ for some constants a and b. **True or false: if the actual covariance of X and Y is 1.5, and you wrongly assume that this covariance is 1.0, then you will overestimate the probability that your portfolio W will suffer a large negative return.** If false, explain why. In answering this question, you may assume that you have correctly specified all the other moments of the joint distribution for X and Y.

D: Let A be the event "rain tomorrow," and let B be the information "the weather app on your phone says it will rain tomorrow." You know that only 15% of all days in Central Texas are rainy. Moreover, you know the track record of your weather app: when it rains, your app gives the correct forecast (and correctly says it will say tomorrow) 90% of the time. When it doesn't rain, your app raises a false alarm (and incorrectly says it will rain tomorrow) 5% of the time. Use this information together with Bayes' rule to calculate the probability of rain tomorrow (A), given that your phone says it will rain tomorrow (B). You need not calculate a final number, but you must simplify your expression as far as possible in terms of the information provided.

Part 2: Essay (40 points)

Imagine that you are trying to build a portfolio of financial assets to save for a new car when you graduate. To keep it simple, let's say you are considering the following choice for allocating your initial capital of \$10,000: 45% of your wealth in an index fund of domestic stocks (ticker SPY), 35% in an index fund of international stocks (ticker INX), and 20% in long-term government bonds (ticker TLT). Naturally, you want to understand the likely risk/return profile of this planned allocation.

Assume that you have access to past daily returns for each asset. By a daily return, I mean the implied interest rate (whether positive or negative) from holding an asset for a single day. More specifically, if $y_{t,i}$ is the value of asset i on day t, and $y_{t-1,i}$ is the value of asset i on day $t-1$, then the return you earned on your holdings of asset i on day t is

$$r_{t,i} = \frac{y_{t,i} - y_{t-1,i}}{y_{t-1,i}}.$$

This is just is the appreciation or depreciation in the value of that asset over the day, expressed as a fraction of the initial value of the asset. Intuitively, you can think of your data set as a large matrix or spreadsheet R: each row is a past day, each column is an asset, and each entry is a return.

Describe in detail how you would use the technique of Monte Carlo simulation, together with bootstrap resampling, to estimate the value at risk (VaR) of your portfolio over a two-week (10-day) horizon. Though you need not write out actual R commands, you must provide a concise description of the logical structure of a computer program that you could write to carry out this analysis, together with the correct equations for calculating important quantities. (We've referred to this as "pseudo-code.") At the same time, pseudo-code alone isn't enough: make sure to address the question of "why?" for each step that must be carried out in the analysis.

Part 3: Interpreting Results (60 points)

In this problem, you will look at weekly data on U.S. railway freight shipments between 1880 and 1886. This was the height of America's so-called Gilded Age, and—like today—a time of rapid technological transformation.

The Joint Executive Committee (or JEC) was a railroad cartel that controlled virtually all east-bound freight shipments from Chicago to the east coast of the United States during the 1880's. All firms involved in the cartel would publicly acknowledge their collusion; this was before the passage of the Sherman Antitrust Act in 1890, which made such anti-competitive arrangements illegal. The JEC would attempt to set rail-shipping prices to maximize the joint profits of all colluding firms. This would be significantly higher than the market price that would prevail if the firms were to compete with each other. Therefore, as you can imagine, individual firms in the cartel faced an incentive to lower their prices in secret, in an attempt to capture more than the market share that they would ordinarily be allocated by the JEC. This made the cartel unstable. In fact, every so often, a price war among the member firms would break out, temporarily ending the collusion before the JEC could regain control over its members and restore the pricing cartel to operational status (see Panel A of Figure 1).

The data set you'll be working with consists of 328 weekly observations of the following variables.

quantity: total tonnage of grain shipped in the week. Shipments of grain accounted for almost three-quarters of all gross-weight tonnage shipped by firms in the JEC.

price: weekly price, in dollars, to ship a dead-weight ton of grain (i.e. not including the weight of the container) by rail for a distance of 100 miles.

cartel: a dummy variable indicating that the railroad cartel was operating during the given week.

season: a categorical variable indicating the time of year, so that season-to-season variation can be modeled. To match the weekly data, the calendar has been divided into 13 periods, each approximately 4 weeks long.

ice: a dummy variable indicating that the Great Lakes have been rendered non-navigable because of ice. If the Great Lakes were open to navigation, then grain could be shipped eastward from Chicago by boat.

The price wars among JEC firms of 1880–86 offer us a natural experiment, where we may gauge the effect that the cartel's operation had on the underlying market dynamics. Specifically, we are are able to test whether the operation of the cartel changed the demand curve for railroad shipping during the period in question. You'll recall from our class discussion (and from your econ class) that a demand curve characterizes the relationship between the price (P) of a certain good or service, and the quantity (Q) of that good or service that consumers are willing to purchase. We will assume a power law: $Q = AP^{\beta}$. The curve has two parameters that may change as a function of outside forces: the multiplicative constant A, and the price elasticity of demand β. Remember that on a log-log scale, this reduces to a linear relationship where

$$\log Q = \log A + \beta \cdot (\log P).$$

Please review the plots and analyses that are summarized over the following pages. Then decide which of the following seven statements are true; which are false; and which are undecidable using the results presented here. Explain your reasoning as concisely as possible (no credit without a correct explanation). If you believe that a statement is undecidable, explain what analysis you'd like to run to decide it. Where relevant, please refer to specific tables and figures in making your case.

(1) The JEC cartel, when it was operational, fundamentally shifted the demand curve for rail shipping up or down, by changing the constant A (holding other relevant factors constant).

(2) The cartel, when it was operational, fundamentally altered the demand curve for rail shipping by changing the price elasticity of demand β (holding other relevant factors constant).

(3) The cartel, when it was operational, did make railroad shipping more expensive for consumers, and therefore affected the quantity demanded. But, holding other relevant factors constant, the cartel's operation did not fundamentally alter the demand curve in either of the ways described in statements 1 or 2.

(4) There are significant seasonal effects in the time series of price (P), adjusting for other relevant factors.

(5) There are significant seasonal effects in the time series of quantity (Q), adjusting for other relevant factors.

(6) The presence of ice on the Great Lakes fundamentally shifted the demand curve for rail shipping up or down, by changing the constant A (holding other relevant factors constant).

(7) The presence of ice on the Great Lakes fundamentally altered the demand curve for rail shipping by changing the price elasticity of demand β (holding other relevant factors constant).

Regressions for price: The following linear regressions were run, all having price as the response.

1. price $\sim$ time
2. price $\sim$ time + ice
3. price $\sim$ time + ice + season
4. price $\sim$ time + ice + cartel
5. price $\sim$ time + ice + season + cartel

The estimated coefficients for these five models are summarized in Table 1. Recall that "season" is a categorical variable with 13 categories, and thus generates 12 dummy variables. These seasonal-dummy coefficients for Models 3 and 5 are shown in Table 3. In addition, three permutation tests were carried out: Model 4 versus Model 2; Model 5 versus Model 3; and Model 5 versus Model 4. The results of these tests are summarized in Figures 2 through 4.

Regressions for log quantity: The following regressions were also run, all having log quantity as the response.

1. log(quantity) $\sim$ log(price)
2. log(quantity) $\sim$ ice + log(price)
3. log(quantity) $\sim$ ice + cartel + log(price)
4. log(quantity) $\sim$ ice + cartel + log(price) + cartel:log(price)
5. log(quantity) $\sim$ season + log(price)
6. log(quantity) $\sim$ season + ice + log(price)
7. log(quantity) $\sim$ season + ice + cartel + log(price)
8. log(quantity) $\sim$ season + ice + cartel + log(price) + cartel:log(price)

Recall that the notation "cartel:log(price)" denotes an interaction term between the cartel dummy variable and the log(price) variable. The estimated coefficients for these eight models are summarized in Table 2. As above, the seasonal-dummy coefficients for Models 5–8 are shown in Table 3. In addition, six permutation tests were carried out: Model 3 versus Model 2; Model 4 versus Model 3; Model 6 versus Model 5; Model 7 versus Model 6; Model 8 versus Model 7; and Model 8 versus Model 4. In each permutation test, the extra variables in the larger model were shuffled. (For example, in testing Model 3 versus Model 2, the cartel variable was shuffled.) The results of these tests are summarized in Figures 5 through 10.

Table 1: Estimated coefficients (along with their standard errors in parentheses) for the five different linear models with price as the response variable. A missing entry in the table indicates that the corresponding variable was excluded from the model in question.

Model	intercept	time	ice	cartel	season	R^2
1	0.3069 (0.0062)	-0.0004 (0.0001)				0.274
2	0.2889 (0.0061)	-0.0004 (0.0001)	0.0464 (0.005)			0.393
3	0.3353 (0.0181)	-0.0003 (0.0001)	0.0438 (0.005)		(see Table 3)	0.415
4	0.2359 (0.0071)	-0.0003 (0.0001)	0.0438 (0.005)	0.0594 (0.008)		0.559
5	0.2593 (0.017)	-0.0003 (0.0001)	0.0136 (0.005)	0.0380 (0.020)	(see Table 3)	0.579

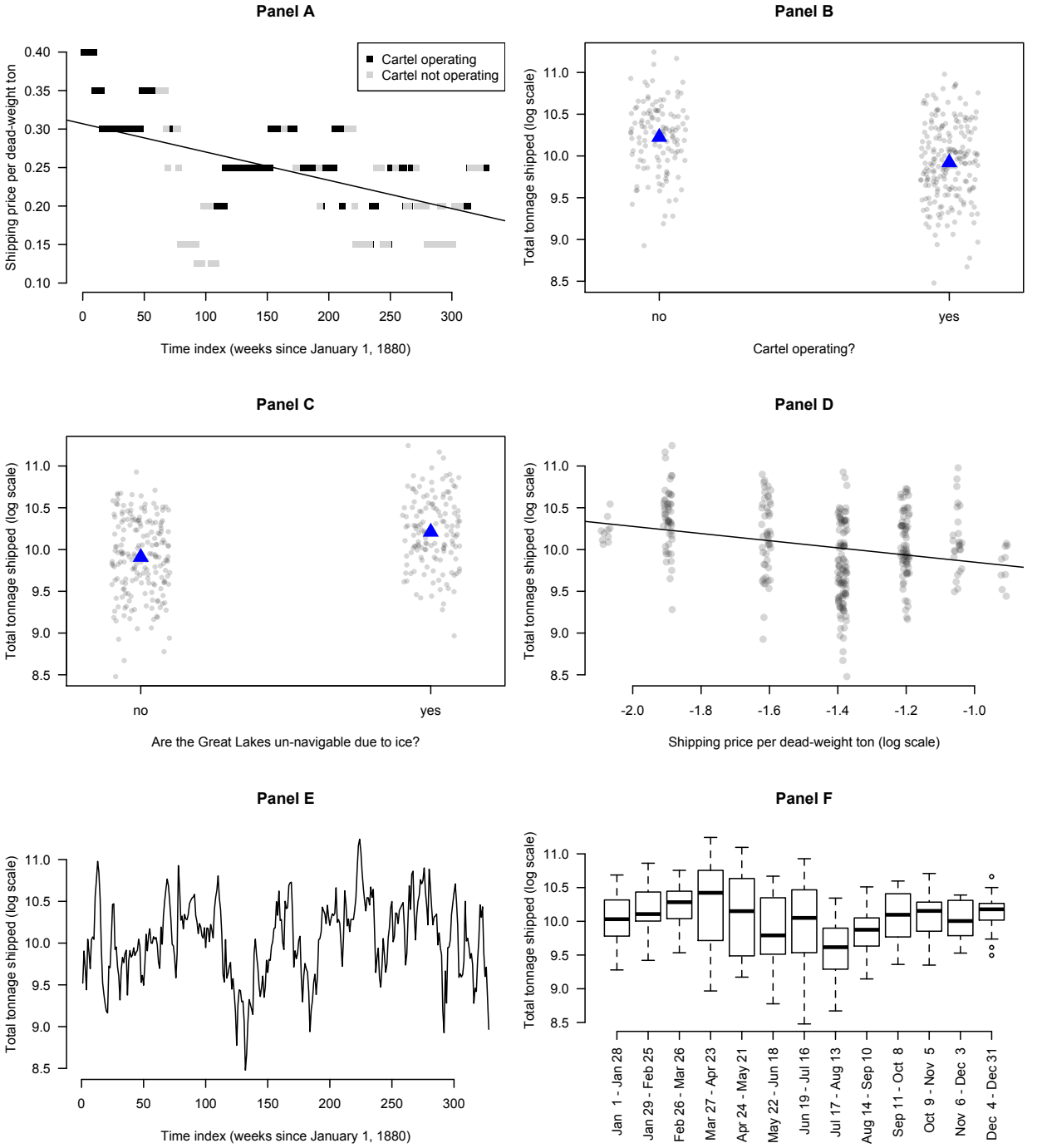

Figure 1: Exploratory analysis for the railroad-cartel data. Panel A: Shipping price per ton charged by JEC firms as it varies over time, with color indicating whether the cartel was operating. Panel B: Total tonnage shipped by members of the rail cartel, stratified by whether the cartel was operating. Panel C: Total tonnage shipped by members of the rail cartel, stratified by whether the Great Lakes were rendered non-navigable by freighters due to the presence of ice. Panel D: Total tonnage shipped by members of the rail cartel versus price, on a log-log scale. Panel E: Total tonnage shipped by members of the rail cartel over time. Panel F: Total tonnage shipped by members of the rail cartel, stratified by thirteen different four-week "seasons" (i.e. months). In panels A and D, the lines are least-squares fits; in Panels B and C, the dark triangles are the group-wise means. In Panels B, C, and D, a small horizontal jitter has been added so that the individual points can be more easily distinguished.

Table 2: Estimated coefficients (along with their standard errors in parentheses) for the eight different linear models with log quantity as the response variable. A missing entry in the table indicates that the corresponding variable was excluded from the model in question.

Model	intercept	log(price)	cartel	cartel : log(price)	ice	season	R^2
1	9.420 (0.127)	-0.428 (0.086)					0.070
2	8.948 (0.122)	-0.634 (0.081)			0.409 (0.047)		0.241
3	9.274 (0.186)	-0.428 (0.086)	-0.142 (0.059)		0.386 (0.048)		0.254
4	9.011 (0.225)	-0.634 (0.129)	0.453 (0.296)	0.413 (0.201)	0.386 (0.048)		0.264
5	9.144 (0.137)	-0.662 (0.084)				(see Table 3)	0.282
6	8.728 (0.178)	-0.639 (0.082)			0.447 (0.119)	(see Table 3)	0.312
7	9.068 (0.230)	-0.488 (0.105)	-0.135 (0.058)		0.409 (0.120)	(see Table 3)	0.324
8	8.680 (0.269)	-0.707 (0.133)	0.646 (0.289)	0.540 (0.199)	0.401 (0.119)	(see Table 3)	0.341

Table 3: Period-by-period detail for the 13 different four-week periods in the data set. The first and fourth rows show the group-wise means for price (P) and log quantity (Q), respectively. The remaining rows show the coefficients for the seasonal dummy variables (with standard errors in parentheses) in each of the regression models for P and log Q where season was included as a categorical predictor.

Date range	1/1 – 1/28	1/29 – 2/25	2/26 – 3/26	3/27 – 4/23	4/24 – 5/21	5/22 – 6/18	6/19 – 7/16	7/17 – 8/13	8/14 – 9/10	9/11 – 10/8	10/9 – 11/5	11/6 – 12/3	12/4 – 12/31
Group mean for P	0.271	0.289	0.288	0.262	0.238	0.233	0.223	0.225	0.223	0.204	0.223	0.241	0.264
coefficient in price model 3		0.019 (0.014)	0.019 (0.014)	-0.005 (0.017)	-0.041 (0.021)	-0.045 (0.021)	-0.054 (0.021)	-0.051 (0.021)	-0.051 (0.021)	-0.069 (0.021)	-0.048 (0.021)	-0.029 (0.021)	0.000 (0.014)
coefficient in price model 5		0.019 (0.012)	0.027 (0.012)	0.005 (0.012)	-0.019 (0.015)	-0.021 (0.018)	-0.03 (0.018)	-0.025 (0.018)	-0.018 (0.018)	-0.036 (0.018)	-0.028 (0.018)	-0.01 (0.018)	-0.008 (0.012)
Group mean for log Q	10.048	10.185	10.229	10.276	10.095	9.864	9.966	9.598	9.846	10.072	10.062	10.003	10.143
coefficient in log Q model 5		0.202 (0.109)	0.247 (0.109)	0.224 (0.108)	-0.021 (0.113)	-0.271 (0.113)	-0.196 (0.114)	-0.554 (0.114)	-0.315 (0.114)	-0.152 (0.115)	-0.104 (0.114)	-0.097 (0.113)	0.096 (0.113)
coefficient in log Q model 6		0.2 (0.106)	0.244 (0.106)	0.288 (0.108)	0.242 (0.131)	0.18 (0.164)	0.255 (0.164)	-0.102 (0.164)	0.136 (0.164)	0.302 (0.165)	0.348 (0.164)	0.352 (0.163)	0.133 (0.111)
coefficient in log Q model 7		0.185 (0.106)	0.21 (0.107)	0.264 (0.107)	0.212 (0.131)	0.149 (0.163)	0.231 (0.164)	-0.134 (0.164)	0.089 (0.164)	0.269 (0.165)	0.33 (0.163)	0.32 (0.163)	0.146 (0.11)
coefficient in log Q model 8		0.222 (0.106)	0.274 (0.109)	0.304 (0.107)	0.236 (0.13)	0.16 (0.162)	0.267 (0.162)	-0.106 (0.162)	0.109 (0.163)	0.274 (0.163)	0.353 (0.162)	0.354 (0.162)	0.172 (0.11)

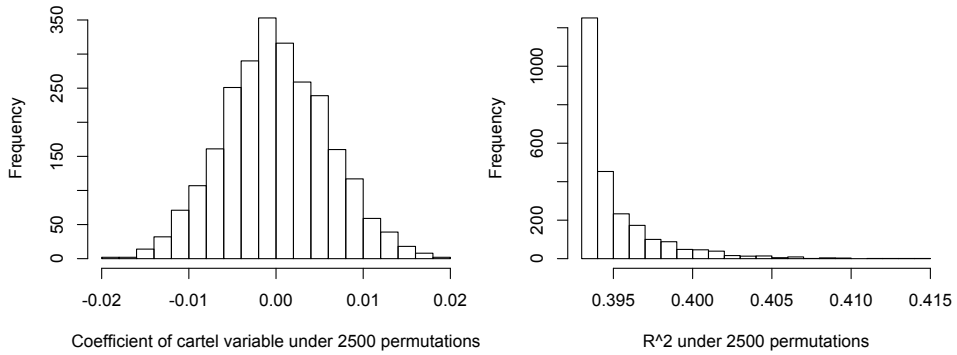

Figure 2: (above). Results of a permutation test of price Model 4 (price versus time, ice, cartel) against Model 2 (price versus time, ice). The left histogram shows the estimated sampling distribution of the cartel variable using 2500 shuffled data sets where the cartel variable has been permuted. At right: R^2 for the same 2500 data sets.

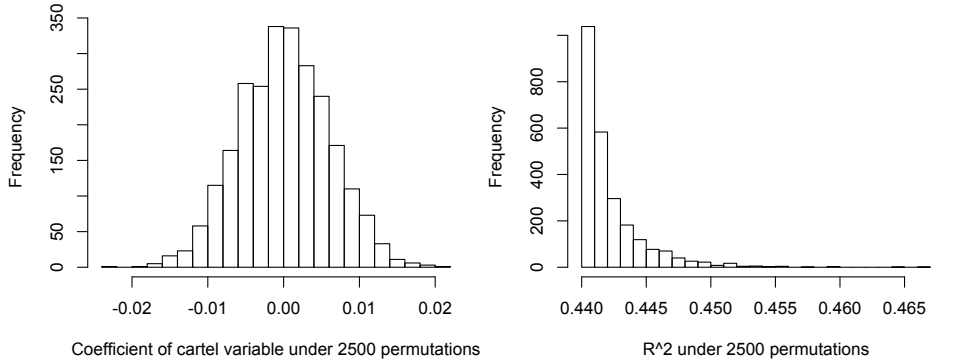

Figure 3: (above). Results of a permutation test of price model 5 (price versus time, ice, season, cartel) against model 3 (price versus time, ice, season). The left histogram shows the estimated sampling distribution of the cartel variable using 2500 shuffled data sets where the cartel variable has been permuted. At right: R^2 for the same 2500 data sets.

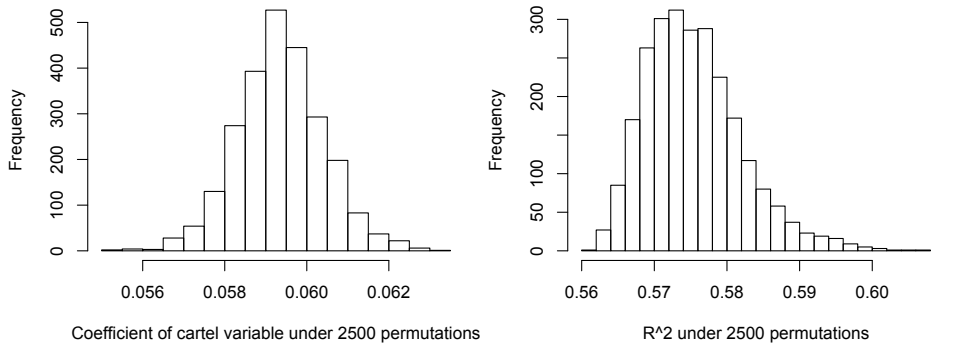

Figure 4: (above). Results of a permutation test of price model 5 (price versus time, ice, season, cartel) against model 4 (price versus time, ice, cartel). The left histogram shows the estimated sampling distribution of the cartel variable using 2500 shuffled data sets where the seasonal dummy variables have been permuted. At right: R^2 for the same 2500 data sets.

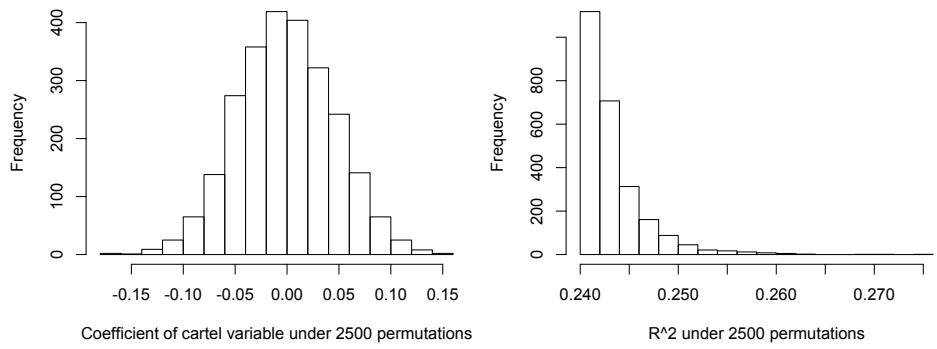

Figure 5: (above). Results of a permutation test of quantity Model 3 (log quantity versus log price, ice, cartel) versus Model 2 (log quantity versus log price, ice). The left histogram shows the estimated sampling distribution of the cartel variable using 2500 shuffled data sets where the cartel variable has been permuted. At right: R^2 for the same 2500 data sets.

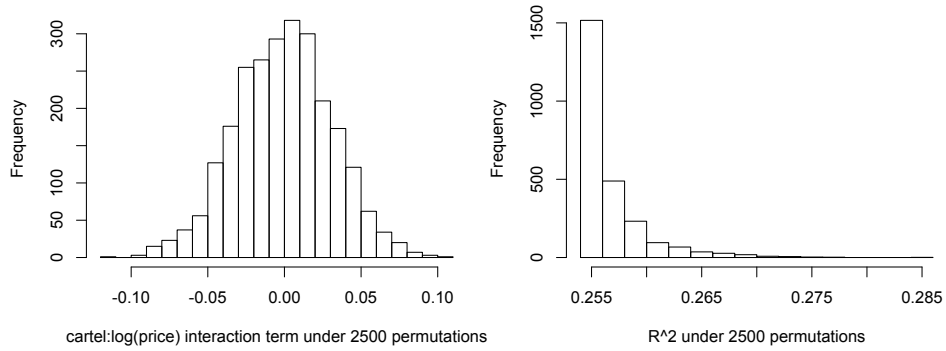

Figure 6: (above). Results of a permutation test of quantity Model 4 (log quantity versus log price, ice, cartel, interaction of cartel and log price) against Model 3 (log quantity versus log price, ice, cartel). The left histogram shows the estimated sampling distribution of the cartel:log(price) interaction term using 2500 shuffled data sets where this interaction variable has been permuted. At right: R^2 for the same 2500 data sets.

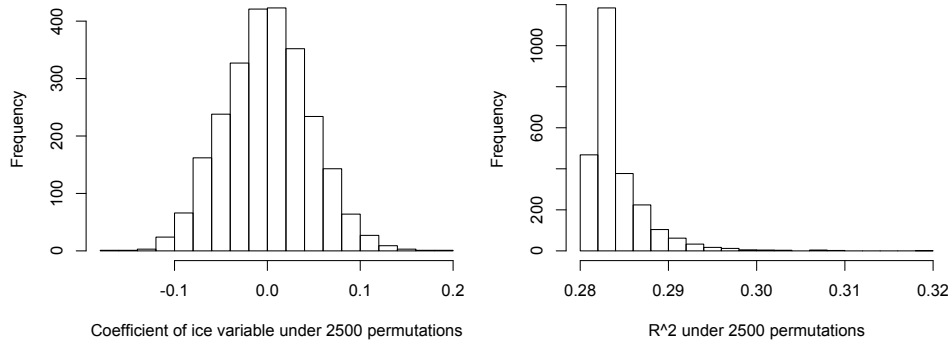

Figure 7: (above). Results of a permutation test of quantity Model 6 (log quantity versus log price, season, ice) against Model 5 (log quantity versus log price, season). The left histogram shows the estimated sampling distribution of the ice variable using 2500 shuffled data sets where the ice variable has been permuted. At right: R^2 for the same 2500 data sets.

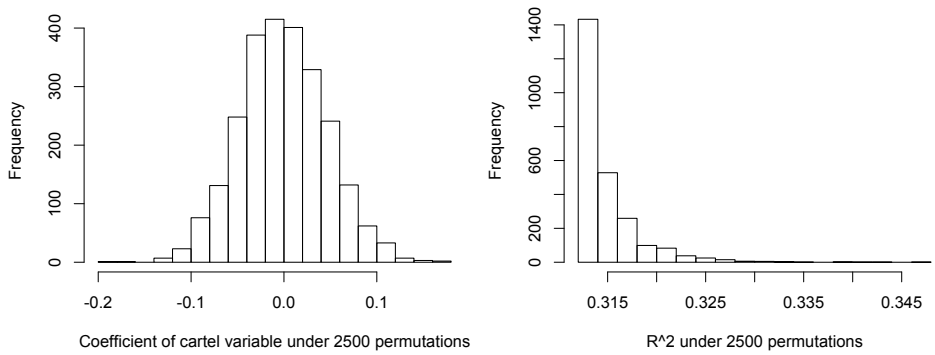

Figure 8: (above). Results of a permutation test of quantity Model 7 (log quantity versus log price, season, ice, cartel) versus Model 6 (log quantity versus log price, season, ice). The left histogram shows the estimated sampling distribution of the cartel variable using 2500 shuffled data sets where the cartel variable has been permuted. At right: R^2 for the same 2500 data sets.

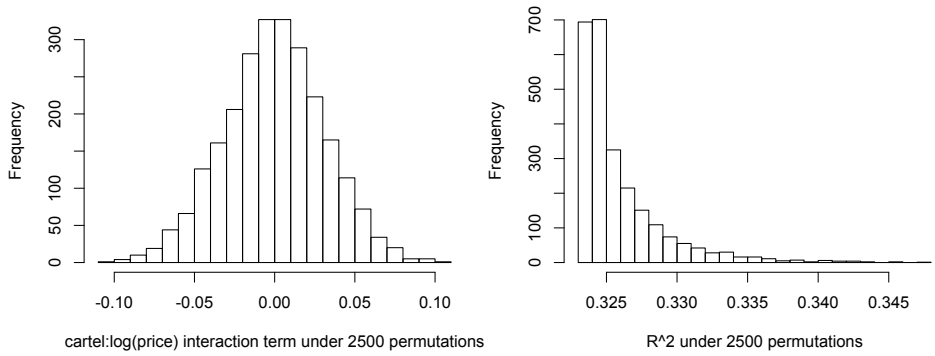

Figure 9: (above). Results of a permutation test of quantity Model 8 (log quantity versus log price, season, ice, cartel, interaction of cartel and log price) against Model 7 (log quantity versus log price, season, ice, cartel). The left histogram shows the estimated sampling distribution of the cartel:log(price) interaction term using 2500 shuffled data sets where this interaction variable has been permuted. At right: R^2 for the same 2500 data sets.

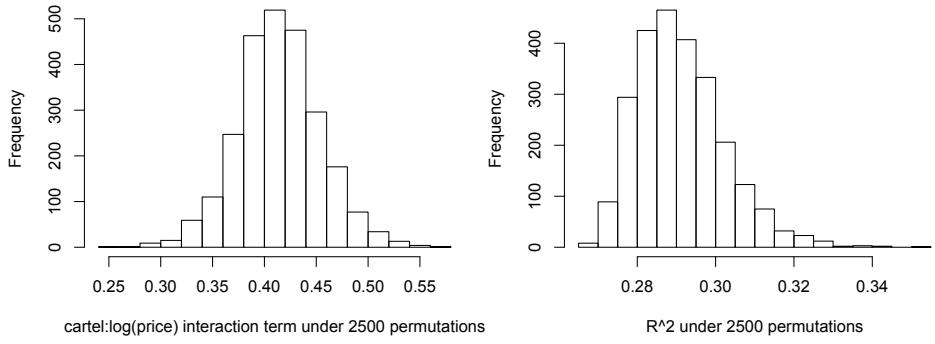

Figure 10: (above). Results of a permutation test of quantity Model 8 (log quantity versus log price, season, ice, cartel, interaction of cartel and log price) against Model 4 (log quantity versus log price, ice, cartel, interaction of cartel and log price). The left histogram shows the estimated sampling distribution of the cartel:log(price) interaction term using 2500 shuffled data sets where the seasonal dummy variables have been permuted. At right: R^2 for the same 2500 data sets.