UNIVERSITY OF TORONTO
Faculty of Arts and Science

APRIL 2015 EXAMINATIONS

ECO220Y1Y, L0101, L0201, L0301, L0401

Duration - 3 hours

Examination Aids: A non-programmable calculator

This exam includes these pages and a separate BUBBLE FORM. Once the exam begins, please detach the 8-page *Attachment* from the end of this exam. The *Attachment* includes your formula sheets, statistical tables (Standard Normal, Student t and F), and a figure and tables for Question 3 in Part 1. The *Attachment* will not be collected. You are responsible for turning in both the BUBBLE FORM and all 18 pages of this exam. You must complete both, including entering your name and student number, before the end of the exam is announced.

This exam has two parts plus the *Attachment*. Part 1 is open-ended questions. Write your answers to Part 1 on these exam papers. Part 2 is multiple choice questions. ***You must record your answers to Part 2 on the BUBBLE FORM. In ALL cases what is (or is not) marked on the BUBBLE FORM is your answer. Marks for Part 2 are based SOLEY on the BUBBLE FORM, which you must complete before the end of the exam is announced.***

Part 1: 3 written questions with varying point values worth a total of 78 points. Write your answers clearly, concisely, and completely below each question. Make sure to show your work and reasoning. Make sure your graphs are fully labeled. A guide for your response ends each question to let you know what is expected: e.g. a quantitative analysis, a graph, and/or sentences. Page 11 gives extra space: use this only if necessary and clearly indicate the question number and make a clear note in the original space directing the grader to it. Unless otherwise specified, you choose the significance level. (If there are no special considerations, you may choose a 5% significance level.)

Part 2: 20 multiple choice questions with point values from 1 to 3 points each for a total of 42 points.

Surname (last name):																			
Given name (first name):																			
Student #:																			

	Q1	Q2	Q3	Part 1 Total	Part 2 Total	Raw Total	Percent Mark
Point Value:	16	16	46	78	42	120	
Points Earned:							

(1) [16pts] Psychology professors at McGill and U of T published "Setting, Elaborating, and Reflecting on Personal Goals Improves Academic Performance" in 2010 in the *Journal of Applied Psychology* http://www.selfauthoring.com/japcomplete.pdf. 25% of students in four-year universities never graduate. The authors randomly divide 85 students in academic difficulty into two groups. The "Goal Group" participated in a web-based 2.5 hour goal-setting program in a lab. The "Control Group" had a similar experience except that instead of a goal-setting program they did things like complete a series of questions to help assess career interests. Since both groups participated in realistic psychological activities, individual participants would not know if they were the in the group the researchers thought would improve (the Goal Group). Both groups competed the lab-based activity in the fall term. The researchers considered GPAs in the winter term right before (GPA1) and the winter term right after (GPA2). For example, if a participant did the lab-based activity in Fall 2014, consider her/his GPA in Winter 2014 (GPA1) and her/his GPA in Winter 2015 (GPA2). See Figure 1, which is taken from the paper.

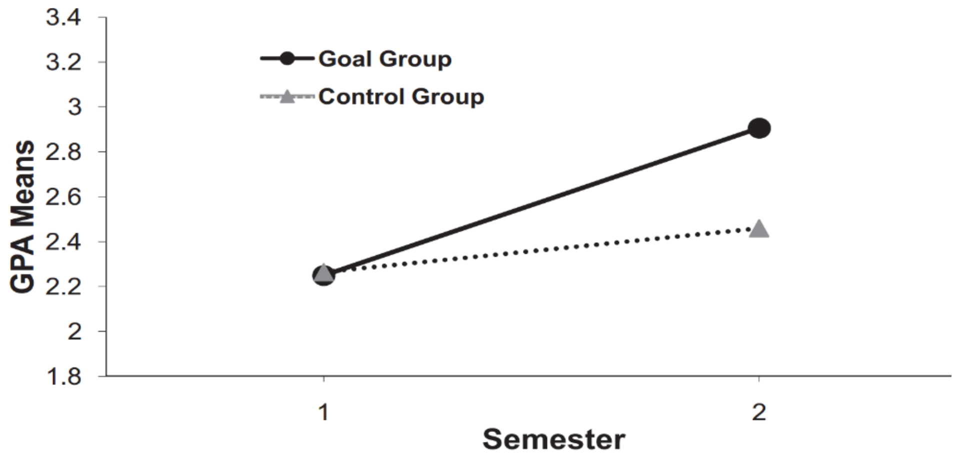


Figure 1. Group differences in grade-point average (GPA) change postintervention.

The table below provides summary statistics:

	Goal Group		**Control Group**	
	mean	s.d.	mean	s.d.
GPA1	2.25	0.93	2.26	0.72
GPA2	2.91	0.65	2.46	1.06
Number of observations	45		40	

(a) [3pts] Why is the mean GPA1 slightly larger for the Control Group compared to the Goal Group? Why is the s.d. of GPA1 bigger for the Goal Group compared to the Control Group? Answer with 1 – 2 sentences making sure to highlight the key course concepts that apply here.

(b) [6pts] Is there sufficient evidence to infer that GPA2 is higher for the Goal Group compared to the Control Group? If so, at which significance levels? Answer with formal hypotheses in standard notation, a quantitative analysis, and 1 sentence. (Note: The correct degrees of freedom to use is 63: skip finding ν yourself.)

(c) [7pts] The authors write "In the Goal Group, the postintervention GPA2 ($M = 2.91, SD = 0.65$) was significantly higher than the baseline GPA1 ($M = 2.25, SD = 0.93$), $t(44) = 4.17$, $p < 0.01$." (It is common notation to write $t(\nu)$ and the P-value as p.) What are the hypotheses? Independent samples or paired data? What is the correlation (r) between GPA1 and GPA2 for the Goal Group? (You can figure it out with the given information.) Answer with formal hypotheses in standard notation and a quantitative analysis.

(2) [16pts] A 2015 NBER working paper "Analyzing the Labor Market Outcomes of Occupational Licensing" (http://www.nber.org/papers/w20961.pdf) investigates if there is a licensing wage premium. In other words, does working in a field that requires a license or certification (e.g. nurse, teacher, truck driver, hair dresser) mean higher wages? The authors have monthly panel data for 25,703 randomly selected workers. Of those, 28.0% are licensed or certified.

(a) [6pts] The analysis includes a simple regression of the natural log of hourly wages on an indicator variable for whether the worker is licensed or certified.

Model (p. 18): $\ln(wage)_{it} = \beta_0 + \beta_1 License_i + \varepsilon_{it}$

Results (p. 41): $b_1 = 0.236, s.e.(b_1) = 0.010$; (Intercept is not reported); $R^2 = 0.033; n = 77{,}294$

Find and *fully interpret* the 99% confidence interval estimator of the OLS slope coefficient. Answer with a quantitative analysis and 1 clear, precise, and complete sentence.

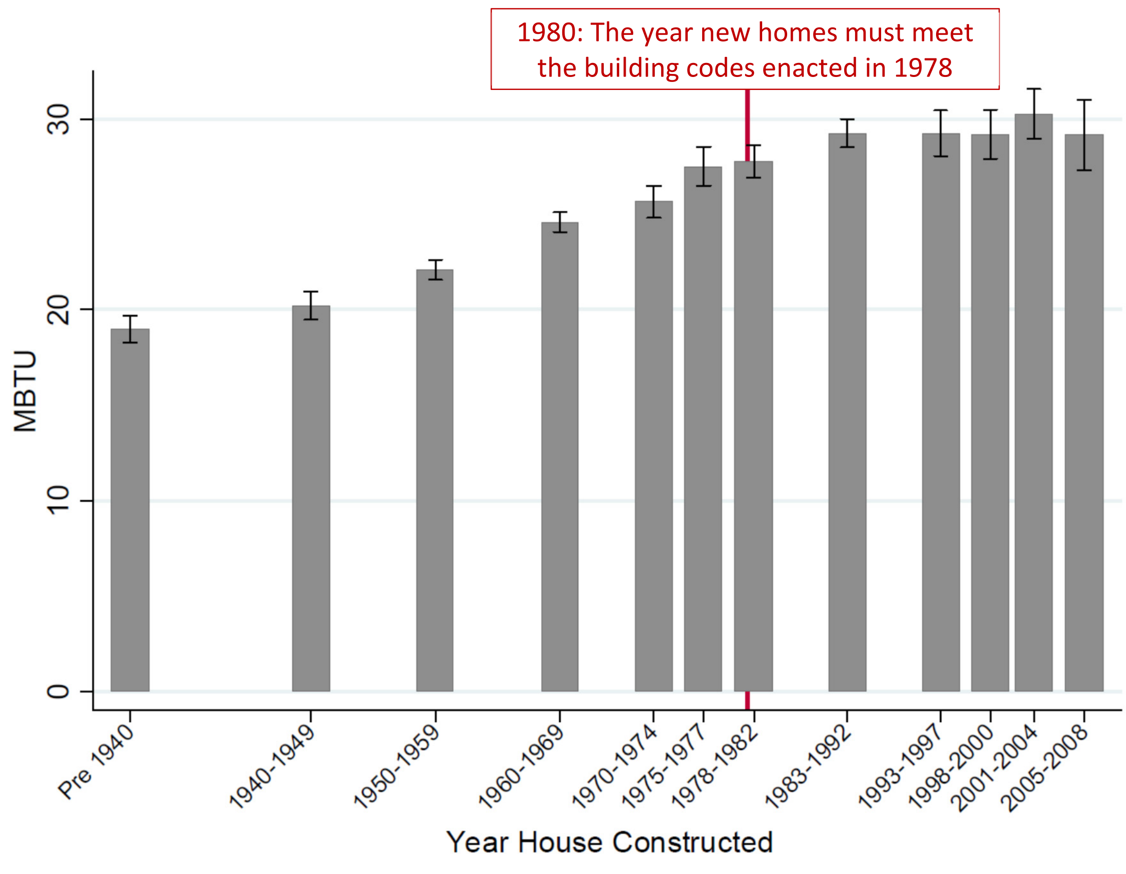
(b) [6pts] A table of descriptive statistics reports that 26.1 percent of the 13,035 males in the sample are licensed and 30.1 percent of the 12,668 females are licensed. Is this difference statistically significant? At which (if any) significance levels? Answer with formal hypotheses in standard notation, a quantitative analysis, and 1 sentence.

(c) [4pts] Consider the hypothesis that being licensed is more important (in term of wages) for females than males. Write down a *multiple regression model specification* that allows testing this hypothesis. Answer with a formal model specification and formal hypotheses in standard notation.

(3) [46pts] A December 2014 NBER working paper "How Much Energy Do Building Energy Codes Really Save? Evidence from California" (http://www.nber.org/papers/w20797) uses multiple regression analysis extensively.

> ***Abstract:*** Construction codes that regulate the energy efficiency of new buildings have been a centerpiece of US environmental policy for 40 years. California enacted the nation's first energy building codes in 1978, and they were projected to reduce residential energy use—and associated pollution—by 80 percent. How effective have the building codes been? I take three approaches to answering that question. First, I compare current electricity use by California homes of different vintages constructed under different standards, controlling for home size, local weather, and tenant characteristics. [Second, ... And third, ...] All three approaches yield the same answer: there is no evidence that homes constructed since California instituted its building energy codes use less electricity today than homes built before the codes came into effect. [Note: Two sentences have been abbreviated with "…".]

The paper focuses on a home's current annual electricity use measured in millions of British thermal units (MBTUs). The *Residential Appliance Saturation Study* (RASS) of 2003 and 2009 provide detailed data on randomly sampled single-family homes without electric heat or hot water in California. The total sample is 16,301 homes (7,417 homes from the 2003 survey and 8,884 homes from the 2009 survey). Because the building codes apply to *new* homes, the year the home was built (its "vintage") is important. Some homes are very old (e.g. built in the 1920's) and some were built just in time for the most recent survey (i.e. built in 2008).



> ***Excerpt (p. 1):*** [The figure] describes the current average annual household electricity use in California, according to when the home was constructed. Homes built recently are not using 80 percent less electricity than homes built before the California standards were first enacted in 1978; they are using more. The comparison is not fair, of course, because homes built more recently are larger, have more occupants, and are in hotter parts of the state. Controlling for those home features, and for the selection of people with high energy demand into recently-built homes, is the objective of this paper.

(a) [4pts] For Table 1 (on page 7 of the *Attachment*), fully interpret the numbers in the second row of results (1.890 and 0.827), the fourth row of results (0.274), and the tenth row of results (0.226 and 0.655). Make sure your interpretations are clear and stated in the easiest way to be understood. Answer with 3 sentences.

(b) [5pts] Recall the popular rule-of-thumb: a regression coefficient is statistically significant at the 5% level if the t test statistic is at least two in absolute value. Conduct a full hypothesis test for the coefficient on "Built 1940s" in Specification (1) (on page 8 of the *Attachment*) and compute its P-value. Would the rule-of-thumb have led to the same conclusion? Answer with formal hypotheses, a quantitative analysis, and 1 – 2 sentences.

Notes: *For the remaining parts of Question 3, use the rule-of-thumb for statistical significance from Part (b). This means you can consider the relevant t statistic and jump to the conclusion without writing out the formal hypotheses and doing full-blown hypothesis tests. Specification (1) – (3) are on page 8 of the Attachment.*

(c) [3pts] In Specification (1), how should the constant term be interpreted? How does it relate to the figure given at the start of this question (which is also reprinted in the *Attachment*)? Answer with 2 sentences.

(d) [6pts] In Specification (1), how should the coefficients for "Built 1950s" and "Built 2005 – 2008" be interpreted? How does each relate to the figure given at the start of this question? Are these statistically significant, economically significant, significant? Answer with 3 – 5 sentences.

(e) [5pts] In Specification (2), fully interpret the coefficient on "ln(Square feet)." In Specification (3), fully interpret the coefficient on "ln(Square feet)." Answer with 1 – 2 sentences remembering to interpret both.

(f) [5pts] In Specification (2), fully interpret the coefficient on "Own home." In Specification (3), fully interpret the coefficient on "Own home." Answer with 1 – 2 sentences remembering to interpret both.

(g) [2pts] In Specification (2), how should the constant term be interpreted? Answer with 1 sentence.

(h) [5pts] In Specification (2), fully interpret the coefficient on "RASS 2009." Does this imply that on average current annual household electricity use increased between 2003 and 2009? Answer with 2 – 3 sentences.

(i) [5pts] For Specification (1), fully interpret the R^2. For Specification (2), fully interpret the R^2. Make sure to comment on its size given the context. Answer with 1 – 2 sentences remembering to interpret both.

(j) [6pts] Why do the coefficient estimates on the home vintage indicators in Specification (2) differ from Specification (1)? Recalling the research question from the abstract ("How effective have the building codes been?"), which conclusions do Specifications (1) – (3) support? Answer with 4 – 6 sentences. (Hint: Re-read both the abstract and the excerpt given at the start of this question and then review Table 2.)

EXTRA SPACE: If you need to use this space, make clear notes directing the grader here.

Part 2: 20 multiple choice questions with point values from 1 to 3 points each for a total of 42 points. Point value for each question shown by [1pt], [2pts], [3pts]. Most questions have choices **(A)** – **(E)**. For questions with fewer choices, the correct answer is ALWAYS one of those offered (e.g. if the choices are **(A)** – **(D)**, then **(E)** is NOT a possible correct answer.)

You must record your answers to Part 2 on the BUBBLE FORM. In ALL cases what is (or is not) marked on the BUBBLE FORM is your answer. Marks for Part 2 are based SOLEY on the BUBBLE FORM, which you must complete before the end of the exam is announced.

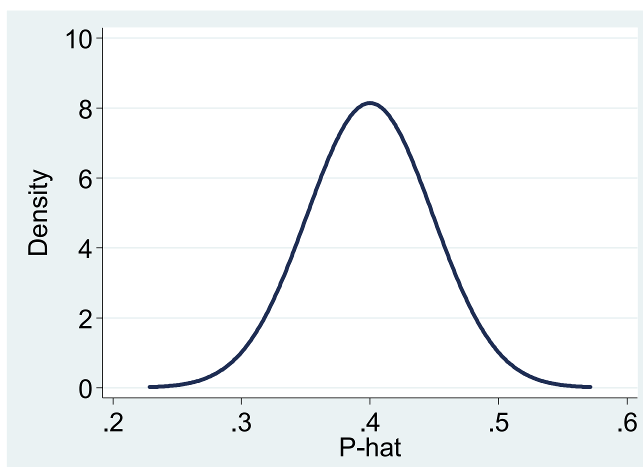
- On the FRONT of the BUBBLE FORM: Print your 9 (or 10) digit student number in the boxes AND darken each number in the corresponding circles; Print your last name and initial in the boxes AND darken each letter in the corresponding circles; Fill in the upper left region of the form. (You may leave the FORM CODE blank.)
- On the BACK of the BUBBLE FORM: Write in your name, sign, and record your answers.
- Use a pencil and make dark solid marks that fill the bubble completely.
- Erase completely any marks you want to change; Crossing out a marked box is incorrect.
- Choose the best answer for each question. If more than one answer is selected that question earns 0 points.
- For questions with numeric answers that require rounding, round your final answer to be consistent with the choices offered. Use standard rounding rules.

REMEMBER, you must record your answers to these 20 multiple-choice questions on the BUBBLE FORM.

▶ **Questions (1) – (2):** Consider this graph of the sampling distribution of a sample proportion.

(1) [2pts] Which sample size matches the graph?

(A) 25
(B) 100
(C) 400
(D) 900
(E) 1,600



(2) [2pts] Compared to the situation in the graph, if the sample size were 5,000, which statement is TRUE?

(A) $P(\hat{P} > 0.30)$ would be much smaller if the sample size were 5,000
(B) $P(\hat{P} < 0.35)$ would be much smaller if the sample size were 5,000
(C) $P(\hat{P} > 0.40)$ would be much smaller if the sample size were 5,000
(D) $P(\hat{P} < 0.40)$ would be much smaller if the sample size were 5,000
(E) All of the above

▶ **<u>Questions (3) – (4):</u>** Recall the publically available data for *all* ON public sector employees with salaries of $100,000 or more. (http://www.fin.gov.on.ca/en/publications/salarydisclosure/pssd/). Below is a STATA summary of 2013 salaries measured in $1000s of dollars.

```
                             salary_2013
-------------------------------------------------------------
      Percentiles      Smallest
 1%     100.2158          100
 5%     100.9888          100
10%     102.1159          100       Obs               95672
25%     105.8357          100       Sum of Wgt.       95672

50%     115.5626                    Mean            127.695
                        Largest     Std. Dev.        38.068
75%     134.0868        772.547
90%     166.2192       903.9706     Variance       1449.173
95%      193.125        915.851     Skewness       4.737556
99%     286.5499           1714     Kurtosis        64.1774
```

(3) [2pts] For one randomly selected ON public sector employee making at least $100,000, what is the probability that that person's salary is over $102,000?

(A) less than 0.6
(B) between 0.6 and 0.7
(C) between 0.7 and 0.8
(D) between 0.8 and 0.9
(E) more than 0.9

(4) [3pts] For a random sample of 180 ON public sector employees making at least $100,000, what is the probability that the sample mean is less than $134,000?

(A) less than 0.75
(B) between 0.75 and 0.90
(C) between 0.90 and 0.95
(D) between 0.95 and 0.99
(E) more than 0.99

▶ **<u>Questions (5) – (10):</u>** Exercises in Chapters 20 and 21 of the textbook explore the relationship between a country's GDP per capita and the regulatory environment. The World Bank has a regulatory quality (RQ) index that captures "perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development." The index (`RQ_index`) is standardized and ranges from approximately -2.5 to 2.5. Higher values mean a higher quality regulatory environment. Among OECD member nations, the RQ index has a mean of 1.27 and a s.d. of 0.44. The World Bank also publishes development indicators including 2012 and 2002 measures of GDP per capita in $1,000s of current US dollars (`gdp_pc_2012` and `gdp_pc_2002`), exports of goods and services as a % of GDP (`exp_pct_gdp`), and Internet users per 100 people (`int_use_100`). For the 34 member nations of the OECD, two regressions are run: Specifications 1 and 2.

```
/* SPECIFICATION 1 */

      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  1,    32) =   24.79
       Model |  8103.96651     1  8103.96651           Prob > F      =  0.0000
    Residual |  10459.2404    32  326.851263           R-squared     =  0.4366
-------------+------------------------------           Adj R-squared =  0.4190
       Total |  18563.2069    33  562.521422           Root MSE      =  18.079

------------------------------------------------------------------------------
 gdp_pc_2012 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    RQ_index |   35.60082   7.149674     4.98   0.000     21.03741    50.16423
       _cons |  -4.769249   9.570234    -0.50   0.622    -24.26318    14.72468
------------------------------------------------------------------------------

/* SPECIFICATION 2 */

      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  4,    29) =   51.10
       Model |  16256.6136     4  4064.15341           Prob > F      =  0.0000
    Residual |  2306.59329    29  79.5376996           R-squared     =  0.8757
-------------+------------------------------           Adj R-squared =  0.8586
       Total |  18563.2069    33  562.521422           Root MSE      =  8.9184

------------------------------------------------------------------------------
 gdp_pc_2012 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    RQ_index |   12.90006   5.485681     2.35   0.026     1.680583    24.11954
 exp_pct_gdp |   .0419379   .0466543     0.90   0.376    -.0534809    .1373566
 int_use_100 |  -.1547858   .1974717    -0.78   0.439    -.5586607    .2490892
 gdp_pc_2002 |   1.585294   .1761316     9.00   0.000     1.225065    1.945524
       _cons |  -.5727771   10.04745    -0.06   0.955    -21.12212    19.97657
------------------------------------------------------------------------------
```

(5) [2pts] For Specification 1, if a country's RQ index is 1, what is the 99% prediction interval for its GDP per capita?

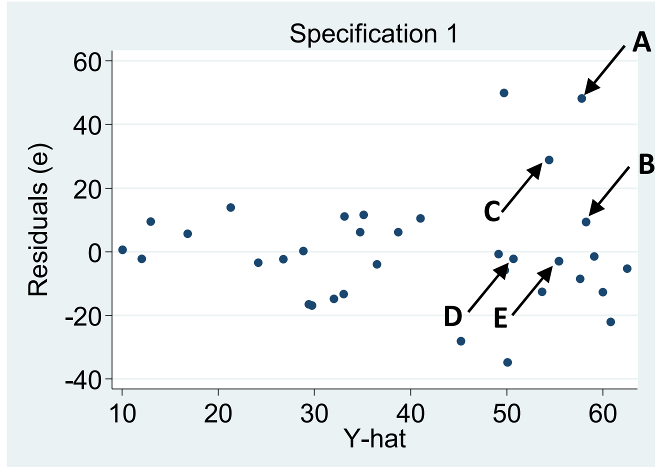
- **(A)** -\$19,669 to \$81,332
- **(B)** -\$14,339 to \$76,002
- **(C)** \$18,955 to \$42,709
- **(D)** \$20,208 to \$41,455
- **(E)** \$22,328 to \$39,335

(6) [2pts] For Specification 1, if 2012 GDP per capita were in dollars (instead of \$1,000s), what would change?

- **(A)** The R^2
- **(B)** The F statistic
- **(C)** The t statistics
- **(D)** The 95% Confidence Interval for the slope coefficient
- **(E)** All of the above

(7) [2pts] For Specification 1, consider the diagnostic plot shown to the right. Canada has a 2012 GDP per capita of $52,400 (current US$) and an RQ index of 1.7. Where is Canada on the graph?

- **(A)** Point A
- **(B)** Point B
- **(C)** Point C
- **(D)** Point D
- **(E)** Point E



(8) [2pts] For Specification 2, approximately what is the critical value for the test of overall statistical significance at a 5% significance level?

- **(A)** 1.96
- **(B)** 2.04
- **(C)** 2.69
- **(D)** 4.17
- **(E)** 51.10

(9) [2pts] For Specification 2, the STATA output says "`Prob > F = 0.0000`." What does this imply?

- **(A)** The multiple regression model overall is highly statistically significant
- **(B)** A larger sample size would have very little effect on the coefficient estimates
- **(C)** It is highly unlikely that the coefficient estimates differ from the true population parameters
- **(D)** The estimated multiple regression model can make highly accurate forecasts of GDP per capita
- **(E)** All of the above

(10) [2pts] For Specification 2, if <u>2002 GDP</u> were in dollars (instead of $1,000s), but everything else was the same as the original specification, which statement is TRUE?

- **(A)** The SSR would be entirely unchanged
- **(B)** The Root MSE would be entirely unchanged
- **(C)** The coefficient on 2002 GDP would be 0.001585294
- **(D)** The standard error for the 2002 GDP coefficient would be 0.0001761316
- **(E)** All of the above

► **<u>Questions (11) – (14):</u>** The next table, which uses U.S. census data for households with a husband and wife, comes from a 2014 NBER working paper "Marry Your Like: Assortative Mating and Income Inequality." (http://www.nber.org/papers/w19829) "Positive assortative mating" means that people tend to marry people with a similar education level. The table has two panels: one for 1960 and another for 2005. A person's highest level of education is categorized as: less than high school (HS-), high school degree (HS), some college (C-), college degree (C), or post-college degree (C+). The rows are husbands' education and columns are wives'.

Contingency Table: Marital Sorting by Education

1960:	*Wife*				
Husband	HS-	HS	C-	C	C+
HS-	0.323	0.138	0.019	0.004	0.001
HS	0.076	0.165	0.028	0.008	0.002
C-	0.018	0.051	0.027	0.008	0.002
C	0.005	0.027	0.019	0.018	0.003
C+	0.003	0.016	0.017	0.016	0.008
2005:	*Wife*				
Husband	HS-	HS	C-	C	C+
HS-	0.039	0.031	0.010	0.003	0.001
HS	0.023	0.192	0.082	0.037	0.012
C-	0.005	0.065	0.088	0.047	0.016
C	0.002	0.030	0.045	0.104	0.037
C+	0.001	0.010	0.018	0.050	0.053

(11) [2pts] In 1960, for households with a husband and wife, which of these scenarios is the *most likely*?

(A) a wife with a high school degree (HS)
(B) a husband with a high school degree (HS)
(C) a husband and wife household where both have a high school degree (HS)
(D) a husband and wife household where both have less than a high school degree (HS-)

(12) [2pts] In 1960, if a husband has a college degree (C), what is the chance his wife has a college degree (C)?

(A) 0.018
(B) 0.054
(C) 0.250
(D) 0.333
(E) 0.351

(13) [3pts] In 2005, if husbands and wives were randomly assigned to each other – in other words, if there is no assortative mating – what number would replace 0.104 (i.e. the cell C, C)?

(A) 0.048
(B) 0.053
(C) 0.106
(D) 0.227
(E) 0.459

(14) [2pts] For 3 randomly selected wives in 2005, what is the chance exactly 1 has a post-college degree?

(A) 0.09
(B) 0.12
(C) 0.18
(D) 0.28
(E) 0.36

▶ **Questions (15) – (17):** Excel output below shows the correlation matrix for short-term interest rates for 6-month, 3-month, and 1-month U.S. Treasury notes as well as regression results for 6M yields using 3M and 1M yields as explanatory variables. What would happen if 3M Treasury were dropped from the regression?

Correlation	6M	3M	1M
6M	1	0.998	0.994
3M	0.998	1	0.998
1M	0.994	0.998	1

Regression Statistics	
Multiple R	0.999
R Square	0.998
Adjusted R Square	0.998
Standard Error	0.078
Observations	164

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	476.345	238.172	38972.592	0.000
Residual	161	0.984	0.006		
Total	163	477.328512			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.062	0.008	7.458	0.000	0.045	0.078
3M Treasury	1.822	0.066	27.565	0.000	1.691	1.952
1M Treasury	-0.817	0.067	-12.149	0.000	-0.950	-0.684

(15) [2pts] If 3M were dropped, the coefficient on 1M Treasury would ___.

(A) stay exactly the same
(B) stay roughly the same
(C) become statistically insignificant
(D) become positive and statistically significant

(16) [1pts] If 3M were dropped, the value 477.328512 (Total, SS) in the *ANOVA* table would ___.

(A) increase
(B) decrease
(C) stay exactly the same

(17) [1pts] If 3M were dropped, the value 0.078 (Standard Error) in the *Regression Statistics* table would ___.

(A) increase
(B) decrease
(C) stay exactly the same

▶ **Questions (18) – (20):** Consider the table and quote excerpted from "Competition and Product Quality in the Supermarket Industry" published in *The Quarterly Journal of Economics* in 2011.

> [The table below] reports summary statistics for supermarkets' characteristics in 1990, before Walmart began entering the grocery industry in earnest. Supermarkets averaged just over 20,000 square feet of selling space. About 1 in 2 supermarkets was part of a chain of 11 or more stores. Stores that would and would not eventually face competition from Walmart were mostly similar across these dimensions, in stark contrast to what would become the footprint of a typical Walmart supercenter. Supercenters average 185,000 square feet, with about a third of that space devoted to groceries.

Firm and Store Characteristics for Walmart and Incumbent Stores

	Grocery Stores, 1990			Walmart Supercenters, 2004
	Full Sample	Eventually Face Walmart Entry	Do Not Face Walmart Entry	
Mean grocery selling space in 1,000s of square feet (s.d. in parentheses)	22.8 (14.5)	23.8 (14.9)	21.1 (13.5)	62.6 (10.5)
Percent affiliated with a chain, where a chain defined as 11 or more stores	57.1	58.2	55.1	100
Observations	28,977	18,218	10,759	1,675

(18) [3pts] Is there a statistically significant difference between grocery stores that eventually face Walmart entry versus those that do not in terms of the mean grocery selling space in 1990?

(A) No
(B) Yes, and the P-value is below 0.001
(C) Yes, and the P-value is between 0.001 and 0.01
(D) Yes, and the P-value is between 0.01 and 0.05
(E) Yes, and the P-value is between 0.05 and 0.10

(19) [3pts] Is there a statistically significant difference between grocery stores that eventually face Walmart entry versus those that do not in terms of the percent of grocery stores affiliated with a chain in 1990?

(A) No
(B) Yes, and the P-value is below 0.001
(C) Yes, and the P-value is between 0.001 and 0.01
(D) Yes, and the P-value is between 0.01 and 0.05
(E) Yes, and the P-value is between 0.05 and 0.10

(20) [2pts] In the excerpt above, what does the statement "Stores that would and would not eventually face competition from Walmart were mostly similar across these dimensions" mean?

(A) There are no statistically significant differences in chain affiliation
(B) There are no statistically significant differences in grocery selling space
(C) There are no sizeable differences in grocery selling space and chain affiliation
(D) All of the above

REMEMBER, you must record your answers to the 20 multiple-choice questions on your BUBBLE FORM.

Sample mean: $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$ **Sample variance:** $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2}{n-1} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n(n-1)}$ **Sample s.d.:** $s = \sqrt{s^2}$

Sample coefficient of variation: $CV = \frac{s}{\bar{X}}$ **Sample covariance:** $s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i}{n-1} - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(n-1)}$

Sample interquartile range: $IQR = Q3 - Q1$ **Sample coefficient of correlation:** $r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{n} z_{x_i} z_{y_i}}{n-1}$

Addition rule: $P(A\ or\ B) = P(A) + P(B) - P(A\ and\ B)$ **Conditional probability:** $P(A|B) = \frac{P(A\ and\ B)}{P(B)}$

Complement rules: $P(A^C) = P(A') = 1 - P(A)$ $P(A^C|B) = P(A'|B) = 1 - P(A|B)$

Multiplication rule: $P(A\ and\ B) = P(A|B)P(B) = P(B|A)P(A)$

Expected value: $E[X] = \mu = \sum_{all\ x} xp(x)$ **Variance:** $V[X] = E[(X-\mu)^2] = \sigma^2 = \sum_{all\ x}(x-\mu)^2 p(x)$

Covariance: $COV[X,Y] = E[(X-\mu_X)(Y-\mu_Y)] = \sigma_{XY} = \sum_{all\ x}\sum_{all\ y}(x-\mu_X)(y-\mu_Y)p(x,y)$

Laws of expected value:

$E[c] = c$

$E[X+c] = E[X] + c$

$E[cX] = cE[X]$

$E[a+bX+cY] = a + bE[X] + cE[Y]$

Laws of variance:

$V[c] = 0$

$V[X+c] = V[X]$

$V[cX] = c^2 V[X]$

$V[a+bX+cY] = b^2V[X] + c^2V[Y] + 2bc * COV[X,Y]$

$V[a+bX+cY] = b^2V[X] + c^2V[Y] + 2bc * SD(X) * SD(Y) * \rho$

where $\rho = CORRELATION[X,Y]$

Laws of covariance:

$COV[X,c] = 0$

$COV[a+bX, c+dY] = bd * COV[X,Y]$

Combinatorial formula: $C_x^n = \frac{n!}{x!(n-x)!}$ **Binomial probability:** $p(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$ for $x = 0,1,2,\ldots,n$

If *X* is Binomial $(X \sim B(n,p))$ **then** $E[X] = np$ **and** $V[X] = np(1-p)$

If *X* is Uniform $(X \sim U[a,b])$ **then** $f(x) = \frac{1}{b-a}$ **and** $E[X] = \frac{a+b}{2}$ **and** $V[X] = \frac{(b-a)^2}{12}$

Sampling distribution of $\bar{X}$:

$\mu_{\bar{X}} = E[\bar{X}] = \mu$

$\sigma_{\bar{X}}^2 = V[\bar{X}] = \frac{\sigma^2}{n}$

$\sigma_{\bar{X}} = SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}$

Sampling distribution of $\hat{P}$:

$\mu_{\hat{P}} = E[\hat{P}] = p$

$\sigma_{\hat{P}}^2 = V[\hat{P}] = \frac{p(1-p)}{n}$

$\sigma_{\hat{P}} = SD[\hat{P}] = \sqrt{\frac{p(1-p)}{n}}$

Sampling distribution of $(\hat{P}_2 - \hat{P}_1)$:

$\mu_{\hat{P}_2-\hat{P}_1} = E[\hat{P}_2 - \hat{P}_1] = p_2 - p_1$

$\sigma_{\hat{P}_2-\hat{P}_1}^2 = V[\hat{P}_2 - \hat{P}_1] = \frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}$

$\sigma_{\hat{P}_2-\hat{P}_1} = SD[\hat{P}_2 - \hat{P}_1] = \sqrt{\frac{p_2(1-p_2)}{n_2} + \frac{p_1(1-p_1)}{n_1}}$

Sampling distribution of $(\bar{X}_1 - \bar{X}_2)$, independent samples:

$\mu_{\bar{X}_1-\bar{X}_2} = E[\bar{X}_1 - \bar{X}_2] = \mu_1 - \mu_2$

$\sigma_{\bar{X}_1-\bar{X}_2}^2 = V[\bar{X}_1 - \bar{X}_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

$\sigma_{\bar{X}_1-\bar{X}_2} = SD[\bar{X}_1 - \bar{X}_2] = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Sampling distribution of $(\bar{X}_d)$, paired $(d = X_1 - X_2)$:

$\mu_{\bar{X}_d} = E[\bar{X}_d] = \mu_1 - \mu_2$

$\sigma_{\bar{X}_d}^2 = V[\bar{X}_d] = \frac{\sigma_d^2}{n} = \frac{\sigma_1^2 + \sigma_2^2 - 2*\rho*\sigma_1*\sigma_2}{n}$

$\sigma_{\bar{X}_d} = SD[\bar{X}_d] = \frac{\sigma_d}{\sqrt{n}} = \sqrt{\frac{\sigma_1^2 + \sigma_2^2 - 2*\rho*\sigma_1*\sigma_2}{n}}$

Inference about a population proportion:

***z* test statistic:** $z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$ **CI estimator:** $\hat{P} \pm z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$

Inference about comparing two population proportions:

***z* test statistic under Null hypothesis of no difference:** $z = \frac{\hat{P}_2 - \hat{P}_1}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_1} + \frac{\bar{P}(1-\bar{P})}{n_2}}}$ **Pooled proportion:** $\bar{P} = \frac{X_1 + X_2}{n_1 + n_2}$

CI estimator: $(\hat{P}_2 - \hat{P}_1) \pm z_{\alpha/2}\sqrt{\frac{\hat{P}_2(1-\hat{P}_2)}{n_2} + \frac{\hat{P}_1(1-\hat{P}_1)}{n_1}}$

Inference about the population mean:

***t* test statistic:** $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ **CI estimator:** $\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ **Degrees of freedom:** $\nu = n - 1$

Inference about a comparing two population means, independent samples, unequal variances:

***t* test statistic:** $t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ **CI estimator:** $(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Degrees of freedom: $\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$

Inference about a comparing two population means, independent samples, assuming equal variances:

***t* test statistic:** $t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$ **CI estimator:** $(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$ **Degrees of freedom:** $\nu = n_1 + n_2 - 2$

Pooled variance: $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

Inference about a comparing two population means, paired data: (n is number of pairs and $d = X_1 - X_2$)

***t* test statistic:** $t = \frac{\bar{d} - \Delta_0}{s_d/\sqrt{n}}$ **CI estimator:** $\bar{X}_d \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$ **Degrees of freedom:** $\nu = n - 1$

SIMPLE REGRESSION:

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ **OLS line:** $\hat{y}_i = b_0 + b_1 x_i$ $b_1 = \frac{s_{xy}}{s_x^2} = r\frac{s_y}{s_x}$ $b_0 = \bar{Y} - b_1 \bar{X}$

Coefficient of determination: $R^2 = (r)^2$ **Residuals:** $e_i = y_i - \hat{y}_i$

Standard deviation of residuals: $s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(e_i - 0)^2}{n-2}}$ **Standard error of slope:** $s.e.(b_1) = s_{b_1} = \frac{s_e}{\sqrt{(n-1)s_x^2}}$

Inference about the population slope:

***t* test statistic:** $t = \frac{b_1 - \beta_{10}}{s.e.(b_1)}$ **CI estimator:** $b_1 \pm t_{\alpha/2} s.e.(b_1)$ **Degrees of freedom:** $\nu = n - 2$

Standard error of slope: $s.e.(b_1) = s_{b_1} = \frac{s_e}{\sqrt{(n-1)s_x^2}}$

Prediction interval for y at given value of x (x_g):

$$\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}} \quad \textbf{or} \quad \hat{y}_{x_g} \pm t_{\alpha/2} \sqrt{\left(s.e.(b_1)\right)^2 (x_g - \bar{X})^2 + \frac{s_e^2}{n} + s_e^2}$$

Degrees of freedom: $\nu = n - 2$

Confidence interval for predicted mean at given value of x (x_g):

$$\hat{y}_{x_g} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_x^2}} \quad \textbf{or} \quad \hat{y}_{x_g} \pm t_{\alpha/2} \sqrt{\left(s.e.(b_1)\right)^2 (x_g - \bar{X})^2 + \frac{s_e^2}{n}}$$

Degrees of freedom: $\nu = n - 2$

SIMPLE & MULTIPLE REGRESSION:

Model: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$

$$SST = \sum_{i=1}^{n} (y_i - \bar{Y})^2 = SSR + SSE \quad SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{Y})^2 \quad SSE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$s_y^2 = \frac{SST}{n-1} \quad MSE = \frac{SSE}{n-k-1} \quad Root\ MSE = \sqrt{\frac{SSE}{n-k-1}} \quad MSR = \frac{SSR}{k}$$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad Adj.R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)} = \left(R^2 - \frac{k}{n-1}\right)\left(\frac{n-1}{n-k-1}\right)$$

Residuals: $e_i = y_i - \hat{y}_i$ **Standard deviation of residuals:** $s_e = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n}(e_i - 0)^2}{n-k-1}}$

Inference about the overall statistical significance of the regression model:

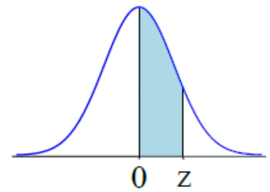
$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{(SST-SSE)/k}{SSE/(n-k-1)} = \frac{SSR/k}{SSE/(n-k-1)} = \frac{MSR}{MSE}$$

Numerator degrees of freedom: $\nu_1 = k$ **Denominator degrees of freedom:** $\nu_2 = n - k - 1$

Inference about the population slope for explanatory variable j:

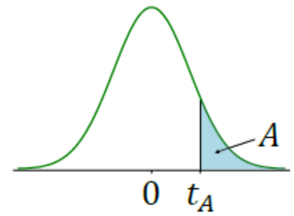
***t* test statistic:** $t = \frac{b_j - \beta_{j0}}{s_{b_j}}$ **CI estimator:** $b_j \pm t_{\alpha/2} s_{b_j}$ **Degrees of freedom:** $\nu = n - k - 1$

Standard error of slope: $s.e.(b_j) = s_{b_j}$ (for multiple regression, must be obtained from technology)



Normal Probabilities:

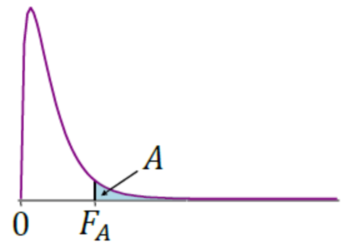
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999



Critical Values of t:

ν	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$	ν	$t_{0.10}$	$t_{0.05}$	$t_{0.025}$	$t_{0.01}$	$t_{0.005}$
1	3.078	6.314	12.706	31.821	63.657	33	1.308	1.692	2.035	2.445	2.733
2	1.886	2.920	4.303	6.965	9.925	34	1.307	1.691	2.032	2.441	2.728
3	1.638	2.353	3.182	4.541	5.841	35	1.306	1.690	2.030	2.438	2.724
4	1.533	2.132	2.776	3.747	4.604	36	1.306	1.688	2.028	2.434	2.719
5	1.476	2.015	2.571	3.365	4.032	37	1.305	1.687	2.026	2.431	2.715
6	1.440	1.943	2.447	3.143	3.707	38	1.304	1.686	2.024	2.429	2.712
7	1.415	1.895	2.365	2.998	3.499	39	1.304	1.685	2.023	2.426	2.708
8	1.397	1.860	2.306	2.896	3.355	40	1.303	1.684	2.021	2.423	2.704
9	1.383	1.833	2.262	2.821	3.250	41	1.303	1.683	2.020	2.421	2.701
10	1.372	1.812	2.228	2.764	3.169	42	1.302	1.682	2.018	2.418	2.698
11	1.363	1.796	2.201	2.718	3.106	43	1.302	1.681	2.017	2.416	2.695
12	1.356	1.782	2.179	2.681	3.055	44	1.301	1.680	2.015	2.414	2.692
13	1.350	1.771	2.160	2.650	3.012	45	1.301	1.679	2.014	2.412	2.690
14	1.345	1.761	2.145	2.624	2.977	46	1.300	1.679	2.013	2.410	2.687
15	1.341	1.753	2.131	2.602	2.947	47	1.300	1.678	2.012	2.408	2.685
16	1.337	1.746	2.120	2.583	2.921	48	1.299	1.677	2.011	2.407	2.682
17	1.333	1.740	2.110	2.567	2.898	49	1.299	1.677	2.010	2.405	2.680
18	1.330	1.734	2.101	2.552	2.878	50	1.299	1.676	2.009	2.403	2.678
19	1.328	1.729	2.093	2.539	2.861	55	1.297	1.673	2.004	2.396	2.668
20	1.325	1.725	2.086	2.528	2.845	60	1.296	1.671	2.000	2.390	2.660
21	1.323	1.721	2.080	2.518	2.831	70	1.294	1.667	1.994	2.381	2.648
22	1.321	1.717	2.074	2.508	2.819	80	1.292	1.664	1.990	2.374	2.639
23	1.319	1.714	2.069	2.500	2.807	90	1.291	1.662	1.987	2.368	2.632
24	1.318	1.711	2.064	2.492	2.797	100	1.290	1.660	1.984	2.364	2.626
25	1.316	1.708	2.060	2.485	2.787	120	1.289	1.658	1.980	2.358	2.617
26	1.315	1.706	2.056	2.479	2.779	140	1.288	1.656	1.977	2.353	2.611
27	1.314	1.703	2.052	2.473	2.771	160	1.287	1.654	1.975	2.350	2.607
28	1.313	1.701	2.048	2.467	2.763	180	1.286	1.653	1.973	2.347	2.603
29	1.311	1.699	2.045	2.462	2.756	200	1.286	1.653	1.972	2.345	2.601
30	1.310	1.697	2.042	2.457	2.750	250	1.285	1.651	1.969	2.341	2.596
31	1.309	1.696	2.040	2.453	2.744	400	1.284	1.649	1.966	2.336	2.588
32	1.309	1.694	2.037	2.449	2.738	∞	1.282	1.645	1.960	2.326	2.576

Degrees of freedom: ν



Critical Values of F: $A = 0.10$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	11	12
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.91	1.89
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79	1.77
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.74	1.71
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.68	1.66
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.63	1.60
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.57	1.55

Critical Values of F: $A = 0.05$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	11	12
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87	1.83
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75

Critical Values of F: $A = 0.01$

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	11	12
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.96	9.89
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34
∞	6.64	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.18

Numerator degrees of freedom: ν_1; Denominator degrees of freedom: ν_2

Figure shows current average annual household electricity use (in MBTU), by when the home was constructed.

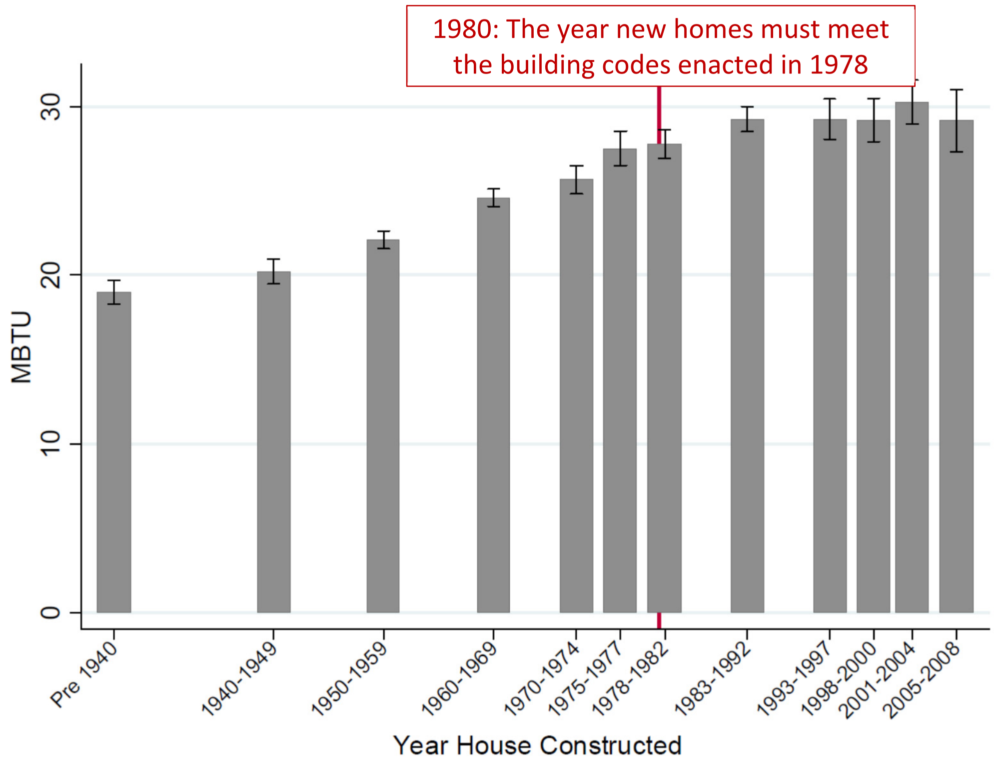


Table 1: Means and Standard Deviations in the RASS

Variable Description	Variable Name	Mean	St. Dev.
1. Current annual HH electricity usage (MBTU) (i.e. in year the home is randomly selected for RASS)	MBTU	25.27	15.98
2. Square feet of residence (1,000s)	Square feet	1.890	0.827
3. Number of bedrooms	Bedrooms	3.279	0.883
4. Indicator for home has an electric stove	Electric stove	0.274	-
5. Indicator for home has an electric oven	Electric oven	0.400	-
6. Indicator for home has been remodeled	Remodeled	0.156	-
7. Years HH has been at this address	Years at address	11.01	6.027
8. Number of residents	Number of residents	2.846	1.490
9. HH income ($1,000s)	HH income ($1,000s)	87.48	57.87
10. Number of residents aged 0 – 5	Number of residents 0 – 5	0.226	0.655
11. Number of residents 65 years old or older	Number of residents 65+	0.504	0.780
12. Indicator for head of HH college graduate	HH head college graduate	0.583	-
13. Indicator for any resident is disabled	Disabled resident	0.102	-
14. Indicator for head of HH is black	HH head black	0.032	-
15. Indicator for head of HH is Latino	HH head Latino	0.126	-
16. Indicator for whether HH owns the home	Own home	0.927	-
17. Indicator for home has central air conditioning	Central AC	0.525	-
18. Indicator for home has room air conditioning	Room AC	0.112	-
19. Number of refrigerators	Refrigerators	1.345	0.534
20. Number of freezers	Freezers	0.260	0.460
21. Indicator for RASS 2009 (versus 2003)	RASS 2009	0.545	-
Observations		16,301	

Table 2: Electricity Consumption in the RASS

	Dependent variable:		
	MBTU	MBTU	ln(MBTU)
Specification:	*(1)*	*(2)*	*(3)*
Home vintage indicators:			
Built 1940s	1.060 (0.519)	0.537 (0.445)	0.094 (0.028)
Built 1950s	3.084 (0.453)	0.585 (0.419)	0.089 (0.024)
Built 1960s	5.565 (0.453)	0.438 (0.432)	0.122 (0.024)
Built 1970 – 1974	6.618 (0.563)	0.692 (0.516)	0.119 (0.027)
Built 1975 – 1977	8.526 (0.638)	0.753 (0.567)	0.157 (0.027)
Built 1978 – 1982	8.604 (0.572)	0.315 (0.524)	0.120 (0.027)
Built 1983 – 1992	10.310 (0.528)	0.562 (0.476)	0.115 (0.025)
Built 1993 – 1997	10.270 (0.710)	-0.248 (0.630)	0.113 (0.028)
Built 1998 – 2000	10.080 (0.754)	-0.751 (0.683)	0.100 (0.030)
Built 2001 – 2004	11.140 (0.760)	-1.344 (0.707)	0.085 (0.030)
Built 2005 – 2008	10.250 (1.018)	-3.904 (0.981)	-0.037 (0.041)
Controls for building characteristics:			
ln(Square feet)	-	9.626 (0.460)	0.306 (0.019)
Bedrooms	-	0.751 (0.186)	0.030 (0.008)
Electric stove	-	-0.858 (0.323)	-0.007 (0.014)
Electric oven	-	2.004 (0.325)	0.068 (0.013)
Remodeled	-	0.726 (0.288)	0.048 (0.013)
Controls for occupant characteristics:			
ln(Years at address)	-	0.810 (0.146)	0.043 (0.008)
ln(Number of residents)	-	5.039 (0.244)	0.251 (0.011)
ln(HH income ($1,000s))	-	2.310 (0.168)	0.082 (0.009)
Number of residents 0 – 5	-	-1.007 (0.181)	-0.044 (0.009)
Number of residents 65+	-	-1.168 (0.143)	-0.039 (0.007)
HH head college graduate	-	-1.325 (0.230)	-0.053 (0.011)
Disabled resident	-	3.348 (0.361)	0.124 (0.016)
HH head black	-	-0.114 (0.507)	0.018 (0.030)
HH head Latino	-	-2.020 (0.278)	-0.075 (0.015)
Own home	-	-3.481 (0.477)	-0.114 (0.020)
Controls for appliances:			
Central AC	-	5.249 (0.238)	0.241 (0.012)
Room AC	-	1.845 (0.311)	0.089 (0.017)
Refrigerators	-	5.724 (0.265)	0.183 (0.010)
Freezers	-	3.042 (0.255)	0.130 (0.011)
Controls for survey year:			
RASS 2009	-	1.980 (0.203)	0.137 (0.011)
Controls for climate:			
13 climate zone dummies	No	Yes	Yes
Constant	19.080 (0.363)	-8.040 (1.006)	1.575 (0.052)
Observations	16,301	16,301	16,301
R^2	0.052	0.360	0.300

Notes: Each column shows a separate regression. Heteroskedastic-consistent standard errors in parentheses. All three regressions are highly statistically significant overall.