Sample Final Exam

STAT-UB.0103 – Statistics for Business Control and Regression Models

Please read this instruction carefully before the exam: The length of the exam is 80 min. The exam is closed book and notes, with the following exception: you are allowed to bring one letter-sized *double-sided* page of notes into the exam. You are also permitted use of a calculator (but smart phone is not allowed).

There are 10 multiple choices questions and one written question. There are in total 100 points. Please circle the choice which best answers each question *in the answer sheet* and and provide answers to written questions *in the answer sheet*. For both multiple choices questions and and written questions, only the answers on the answer sheet will be graded. *Answers not on the answer sheet will NOT be graded.* But you *have to turn in the whole booklet. The exam will be invalid if any page from this booklet is missing.*

Additionally, it is your responsibility to make sure that the Teaching Fellow (TF) can clearly read your answer (if your circle for the multiple choice question is not clear enough or your handwriting is unreadable to TF, TF has the right to claim that this answer is wrong). Therefore, if you do need to modify your answer, make sure TF can easily recognize your final answer.

NYU Stern Honor Code:

I will not lie, cheat or steal to gain an academic advantage, or tolerate those who do.

Signature: _____ Date: _____

Printed Name: _____ Net ID: _____

Leave this page blank

Sample Final Exam

STAT-UB.0103 – Statistics for Business Control and Regression Models

1) (A) (B) (C) (D) (E)

2) (A) (B) (C) (D) (E)

3) (A) (B) (C) (D) (E)

4) (A) (B) (C) (D) (E)

5) (A) (B) (C) (D) (E)

6) (A) (B) (C) (D) (E)

7) (A) (B) (C) (D) (E)

8) (A) (B) (C) (D) (E)

Leave this page blank

Answer to Question 9.

(a)

(b)

(c)

(d)

Leave this page blank

Answer to Question 10.

(a)

(b)

(c)

Leave this page blank

Answer to Question 10 (continue)

(d)

(e)

(f)

Leave this page blank

Multiple Choice

1. (5 points) Suppose you observe a sample data of size n, $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$. The sample variance of x is defined as $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$; the sample variance of y is defined as $\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$; the sample covariance of x and y is defined as $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$, where $\bar{x}$ and $\bar{y}$ are the sample means. From an observed sample data of size $n = 101$, you calculated that the sample covariance of x and y is 6, sample variance of x is 16, and sample variance of y is 4. You are working with the simple regression model $y = \beta_0 + \beta_1 x + \epsilon$. Given x, an estimate of the variance of y is:

   A. $s^2 \approx 3.98$

   B. $s^2 \approx -2.36$

   C. $s^2 \approx 2.36$

   D. $s^2 \approx 1.77$

   E. None of the above

2. (5 points) Consider a multiple linear regression model with three predictive variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon.$$

   Suppose the sample is of size n. To test the following hypothesis

$$H_0 : \beta_2 = 0, \quad \text{versus } H_a : \beta_2 \neq 0,$$

   which statistic and distribution should you use to construct a rejection region or p-value?

   A. The sampling distribution of $\frac{\beta_2}{s_{\beta_2}} \sim t_{n-4}$

   B. The sampling distribution of $\frac{\hat{\beta}_2}{s_{\hat{\beta}_2}} \sim t_{n-4}$

   C. The sampling distribution of $\frac{\beta_2 - 0.7}{s_{\beta_2}} \sim t_{n-4}$

   D. The sampling distribution of $\frac{\hat{\beta}_2 - 0.7}{s_{\hat{\beta}_2}} \sim t_{n-4}$

   E. None of the above

3. (5 points) In multiple linear regression based on a very large sample (very large n), suppose that the t-statistic for one of the regression coefficients is 2.4. Then the two-tailed p-value corresponding to the given coefficient is (note that by the z-table, $\mathrm{P}(0 < Z < 2.4) = .4918$ where Z stands for standard Normal random variable):

   A. .9918

   B. .0164

   C. .9836

   D. .0082

   E. None of the Above

4. (5 points) Consider two simple linear regression data sets, Set1 and Set2. If the sum of squared error SSE for Set2 is smaller than that for Set1, then

   A. the Rsquare for Set2 must be larger than the Rsquare for Set1

   B. the Rsquare for Set2 must be smaller than the Rsquare for Set1

   C. the Rsquare for Set2 must be the same as the Rsquare for Set1

   D. Cannot be determined

5. (5 points) Sales data for a line of high-performance sports cars were recently analyzed to predict the selling price based on various features. The response variable y is the selling price (measured in thousands of dollars), the first predictor x_1 is the engine power (measured in horsepower), and the second predictor x_2 is the interior leather quality (rated on a scale from 1 to 10). The fitted multiple linear regression model is $y = 20 + 2.5x_1 + 8x_2$. Interpret the estimated slope for interior leather quality (that is, $\hat{\beta}_2$)

    A. For every 1-point increase in leather quality, the expected price increases by 8 thousand dollars

    B. For every 1 thousand dollars increase in price, the expected leather quality increases by 8 points

    C. For every 1-point increase in leather quality, the expected price increases by 8 thousand dollars, holding engine power constant

    D. For every 1 thousand dollars increase in price, the expected leather quality increases by 8 points, holding engine power constant

    E. None of the above

6. (5 points) In a linear regression with four predictive variables, if the sample size is 40, the total sum of squares SS_{yy} is 20, the regression sum of squares SSR is 5, what is the estimated standard error s:

    A. .3947

    B. .6283

    C. .7500

    D. .2500

    E. None of the Above

7. (5 points) Consider the same simple linear regression and regression report as in Question 6. What is the estimated standard error s?

   A. 0.655

   B. 0.628

   C. 0.429

   D. 0.395

   E. None of the above.

8. (5 points) Consider the response variable y and two possible predictors of y – variables p and q. Plot A shows the regression of y on p and plot B shows the regression of y on q. Let $\hat{\epsilon}_p$ denote the residuals from the regression of y on p and let $\hat{\epsilon}_q$ denote the residuals from the regression of y on q. Which of the following conclusions is true?

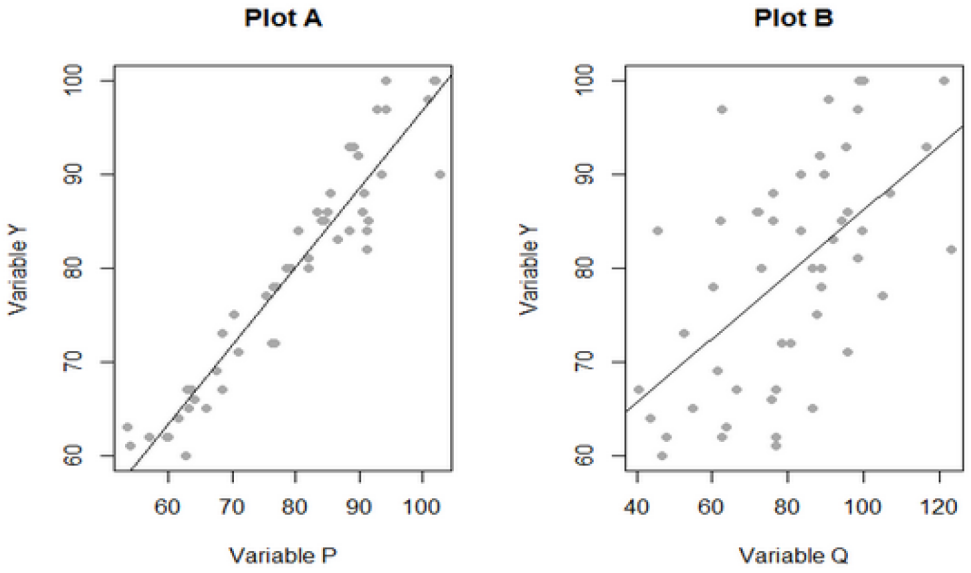
   A. p is the better predictor of y since $Var(\hat{\epsilon}_p)$ is greater than $Var(\hat{\epsilon}_q)$.

   B. p is the better predictor of y since $Var(\hat{\epsilon}_p)$ is less than $Var(\hat{\epsilon}_q)$.

   C. q is the better predictor of y since the regression of y on q has the greater r^2.

   D. q is the better predictor of y since the regression of y on q has the smaller r^2.

   E. None of the above.



Written Problems

Important note: please write down your answers in the answer sheet (not here!). In addition to the final numerical solution in the answer sheet, please provide all the detailed intermediate steps and formulas in the answer sheet.

9. (50 points) In order to investigate the feasibility of starting a Sunday edition for a large metropolitan newspaper, information was obtained from a sample of newspapers concerning their Daily and Sunday circulations (both **in thousands**). In the Minitab output below **Daily** denotes daily circulation and **Sunday,** a dependent variable**,** denotes Sunday circulation. The basic descriptive statistics, scatter plot and the regression analysis output from the Minitab are shown below (some numbers are concealed and denoted by ①,②,. . . ,etc).

```
Descriptive Statistics: Sunday, Daily

Variable   N   Mean  StDev  Minimum  Maximum
Sunday    32  614.4  375.9    262.0   1762.0
Daily     32  441.0  273.7    133.2   1209.2
```

```
Regression Analysis: Sunday versus Daily

The regression equation is
Sunday = 30.67 + 1.324 Daily


S = ②   R-Sq =③


Analysis of Variance

Source      DF       SS
Regression   1   4070375
Error       30    310843
Total       31      ④

Coefficients

Term        Coef  SE Coef  T-Value  P-Value
Constant   30.67     34.5     0.89    0.381
Daily      1.324   0.0668      ⑤      ⑥
```

(a) Compute the t-value in ⑤ (round to the 2nd decimal place) using the MINITAB output. Is the regression of Sunday circulation on daily circulation ***statistically significant*** at $\alpha = 0.05$? Note you need to state (1) the hypothesis test, (2) rejection region and (3) your conclusion.

(b) Compute the total sum of squares $SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ in ④ using the MINITAB output.

(c) Compute the estimated standard error of the regression model s in ② using the MINITAB output (round to the 3rd decimal place).

(d) Compute coefficient of determination (r^2) in ① using the result from the previous question (round to the 3rd decimal place). And interpret the value of r^2 you obtained.

(e) One is interested in the expected Sunday circulation when daily circulation is 400 (thousand). Calculate the fitted value and construct 95% confidence interval (round to the 1st decimal place).

(f) Construct 95% prediction interval for sales when daily circulation is 400 (thousand) (round to the 1st decimal place).

10. (50 points) A real estate broker constructed a multiple linear regression model for house prices in his area using as explanatory variables size of a house, age of a house, lotsize and source of energy used to heat a house. He introduced the following variables:

- Price: sale price of a house (in $1000)
- Size: size of a house (in 100 sq ft)
- Age: age of a house (in years)
- Lotsize: size of a lot (in 1000 sq ft)

There are three sources of energy: electricity, heating oil and natural gas. He coded this information as follows

- Gas = 1 if a house is heated by natural gas and Gas = 0 if not
- Oil = 1 if a house is heated by heating oil and Oil = 0 if not

Answer the questions below based on the Minitab output on the next page (some numbers are concealed and denoted by ①,②,. . . ,etc).

```
Regression Analysis: Price versus hsize, oil, gas, age, lotsize

Analysis of Variance

Source      DF       SS      MS  F-Value  P-Value
Regression  ①      ④     ⑥    ⑧    0.000
Error       9        ⑤     ⑦
Total       ③     6230.24
```

Model Summary

S	R-sq
4.71154	⑨

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-5.5	13.7	-0.40	0.696
hsize	4.143	0.517	8.01	0.000
oil	-10.79	3.07	-3.52	0.007
gas	2.67	4.23	0.63	0.544
age	0.143	0.610	0.23	0.820
lotsize	3.137	0.944	3.32	0.009

Regression Equation

Price = -5.5 +4.143 hsize -10.79 oil +2.67 gas

+0.143 age +3.137 lotsize

(a) What is the degree of freedom for sum of squares due to regression (SSR) in ① ?

(b) What is sum of squared errors SSE in ⑤ (round to the 2nd decimal place)?

(c) What is sum of squares due to regression (SSR) in ④ (round to the 2nd decimal place)?

(d) What is the F-value in ⑧ (round to the 2nd decimal place)?

(e) Calculate r^2 in ⑨ and interpret it.

(f) Is this **overall** multiple linear regression model **statistically significant** at $\alpha = 0.05$ (state the hypothesis, rejection rule and your conclusion)

(g) Identify and Interpret the estimated parameter of Gas and the estimated parameter of Oil.

(h) For houses of the same characteristics, which energy source gives, on average, the highest estimated sale price? Which gives the lowest

price?

(i) Can you conclude at 1% level of significance that, all else equal, houses heated by oil sell for *less than* those heated by electricity (state hypothesis, rejection region and conclusion)?

(j) Construct a 95% confidence interval for the difference in average prices of the houses heated by gas and those heated by electricity holding all other predictors constant.

α	z_α
0.01	2.33
0.05	1.645
0.025	1.960
0.005	2.575

Table entry for p and C is the critical value t^* with probability p lying to its right and probability C lying between $-t^*$ and t^*.

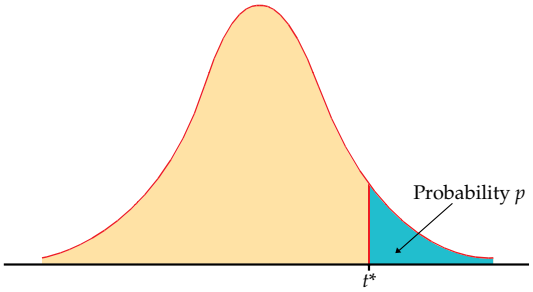


TABLE D

t distribution critical values

	Upper-tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	0.675	0.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
z^*	0.674	0.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											