

**Cornell University**  
**New York State School of Industrial and Labor Relations**  
**Department of Social Statistics**  
**Spring Term 2012**  
**ILRST 5110: Statistical Methods for the Social Sciences II**  
**ILRST 2110: Statistical Methods for the Social Sciences II**

**TEST #1**

**INSTRUCTIONS:** Please answer all questions and write your answers neatly on the test paper. For each question worth 1 point, a short answer is required that can typically be read directly from the given computer printout or other information provided; for these questions, simply report the required answer without further explanation. For questions worth 2 or more points, additional calculation or reasoning is necessary, and you are required to show how you derived your answer by providing a relevant calculation or explanation; for these questions, please make your answers as brief as possible subject to clarity, and be sure to include any relevant computer printout (which you should append at the end of the test paper) if you indeed used the computer in your solution. The test has a total of 105 points.

Remember that the solutions you submit should reflect your own reasoning and not be copied from elsewhere.

A "hard-copy" of your solutions is due in class on Tuesday, April 3.

**1. Write your name here:**

**2. Write your section day here:**

**3. Write your section time here:**

**PART A. [33 points]** A large multinational corporation has hired a team of HRM consultants to evaluate the effectiveness of the human resource management practices at one of its divisions. The consultants are famous for having developed a proprietary scale (ENG) that measures employee engagement on a scale of one to fifty. To assess the level of engagement among the employees at the division, a number of workers are sampled at random and administered an extensive questionnaire. Based on the responses to the questionnaire, each of the employees is assigned an ENG score by the consultants. The following printout shows one part of the statistical analysis performed by the consultants on the data they obtained:

### One-Sample T: ENG

Test of mu = 36.4 vs not = 36.4

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
ENG	100	35.200	6.500	0.650	(33.910, 36.490)	-1.85	0.068

In the questions that follow, let  $\mu$  be the mean of the population of ENG scores for workers at the division, and let  $\sigma^2$  be the population variance, so  $\sigma$  denotes the population standard deviation. Similarly, let  $\bar{y}$  denote the mean of the sample of ENG scores, and let  $s^2$  denote the sample variance, so  $s$  denotes the sample standard deviation.

1. [1 point] What is the sample size, i.e., how many employees were chosen at random and administered the questionnaire?
2. [1 point] What is the value of the sample mean  $\bar{y}$ ?
3. [1 point] Give an estimate of the population mean  $\mu$ .
4. [1 point] What is the value of the sample standard deviation  $s$ ?
5. [2 points] What is the value of the sample variance  $s^2$ ?

6. [1 point] Give an estimate of the population standard deviation  $\sigma$ .
  
7. [1 point] Give an expression for the standard deviation of  $\bar{y}$  in terms of the population standard deviation  $\sigma$ .
  
8. [1 point] Give an estimate of the standard deviation of  $\bar{y}$ .
  
9. [1 point] Give a 95% confidence interval for the population mean  $\mu$ .
  
10. [3 points] Construct a 90% confidence interval for  $\mu$ .
  
11. [1 point] State the null hypothesis being tested in the printout.
  
12. [1 point] State the alternative hypothesis being used for the test in the printout.
  
13. [1 point] Give the value of the test statistic for the hypothesis test performed in the printout.

14. [1 point] Give the  $p$ -value for the hypothesis test performed in the printout.

15. [1 point] What distribution is being used to compute the  $p$ -value for the hypothesis test performed in the printout, i.e., what is the assumed distribution of the test statistic under the null hypothesis?

16. [2 points] For the test performed in the printout, would the null hypothesis be rejected in favor of the alternative hypothesis in a test at the 5% level?

17. [2 points] Give the  $p$ -value for testing the null hypothesis  $H_0 : \mu = 36.4$  against the alternative hypothesis  $H_a : \mu < 36.4$ .

18. [2 points] Would the null hypothesis  $H_0 : \mu = 36.4$  be rejected in favor of the alternative hypothesis  $H_a : \mu < 36.4$  in a test at the 5% level?

19. [2 points] Give the  $p$ -value for testing the null hypothesis  $H_0 : \mu = 36.4$  against the alternative hypothesis  $H_a : \mu > 36.4$ .

20. [2 points] Would the null hypothesis  $H_0 : \mu = 36.4$  be rejected in favor of the alternative hypothesis  $H_a : \mu > 36.4$  in a test at the 5% level? (Justify your answer briefly.)

21. [2 points] During a previous study commissioned by the corporation, the consultants determined by a complete census that the population mean of the ENG scores for the employees in a different, much smaller division which uses radically different HRM practices is 33.7. The corporation is interested in knowing if the population mean of the ENG scores for the division currently under study is greater than that for the division that was investigated previously. Consequently, the consultants decide to test the null hypothesis  $H_0 : \mu = 33.7$ . What is the value of the test statistic for testing this null hypothesis?

22. [3 points] The consultants decide to test the null hypothesis  $H_0 : \mu = 33.7$  against the alternative  $H_a : \mu > 33.7$ . Would the null hypothesis be rejected in favor of the alternative hypothesis in a test at the 1% level?

**PART B. [32 points]** In addition to measuring the engagement of each of the sampled employees, the consultants also measured the amount of supervisor support that each of the sampled employees perceive they receive. The amount of supervisor support is measured by a SUP score on a scale of one to fifty. Using the ENG and SUP scores, the consultants performed the following simple linear regression analysis:

## Regression Analysis: ENG versus SUP

The regression equation is  
ENG = 14.2 + 0.600 SUP

Predictor	Coef	SE Coef	T	P
Constant	14.200	2.646	5.37	0.000
SUP	0.60000	0.07420	8.09	0.000

S = 5.05966    R-Sq = 40.0%    R-Sq(adj) = 39.4%

### Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1			65.39	0.000
Residual Error	98	2508.8	25.6		
Total	99	4182.8			

### Predicted Values for New Observations

New Obs	Fit	SE Fit	95% CI	95% PI
1	39.400	0.725	(37.961, 40.839)	(29.256, 49.543)

### Values of Predictors for New Observations

New Obs	SUP
1	42.0

In the questions that follow, let the population regression line be denoted by  $E(y|x) = \beta_0 + \beta_1 x$ , where the response variable  $y$  is ENG and the predictor variable  $x$  is SUP. Recall that the simple linear regression model assumes constant variance, i.e.,  $Var(y|x) = \sigma^2$ .

1. [1 point] Give the fitted least-squares regression line.
2. [1 point] Give the value of  $b_0$ , the estimate of  $\beta_0$ , the population intercept.
3. [1 point] Give the value of  $b_1$ , the estimate of  $\beta_1$ , the population slope.

4. [1 point] Estimate the change in the mean ENG score per one-point increase in SUP score.
5. [2 points] Estimate the mean of the population of ENG scores for employees who have an SUP score of 35.
6. [1 point] Estimate the standard deviation of the population of ENG scores for employees who have an SUP score of 35.
7. [1 point] Estimate the standard deviation of  $b_0$ .
8. [1 point] Give the value of the test statistic for testing the null hypothesis  $H_0 : \beta_1 = 0$ .
9. [1 point] Which distribution is used to calculate the  $p$ -value for testing  $H_0 : \beta_1 = 0$ ?
10. [2 points] Would you reject the null hypothesis  $H_0 : \beta_1 = 0$  in favor of the alternative hypothesis  $H_a : \beta_1 \neq 0$  in a test at the 1% level?

11. [2 points] Is there evidence that the SUP scores are useful for explaining the variability in the ENG scores?

12. [1 point] What is the amount of variability in the ENG scores that is left unexplained after the SUP scores are taken into account through a linear relationship?

13. [2 points] What is the amount of variability in the ENG scores that is explained by taking the SUP scores into account through a linear relationship?

14. [1 point] What is the proportion of variability in the ENG scores that is explained by taking the SUP scores into account through a linear relationship?

15. [1 point] What is the coefficient of determination for the regression?

16. [2 points] Give the value of the correlation coefficient between the ENG and SUP scores.

17. [3 points] Observation 9 in the data set has  $SUP = 39.3$  and  $ENG = 35.6$ . What are the fitted value and the residual for this observation?



18. [3 points] Construct a 95% confidence interval for  $\beta_1$ .

19. [1 point] Give a 95% confidence interval for the mean ENG score of employees who have an SUP score of 42.

20. [2 points] What is the value of the test statistic for testing  $H_0 : E(y | x = 42) = 37$ , i.e., for testing the null hypothesis that the mean ENG score is 37 for employees having an SUP score of 42?

21. [1 point] Suppose that a randomly chosen employee has an SUP score of 42. Give a point prediction of the employee's ENG score.

22. [1 point] Give a 95% prediction interval for the ENG score of a randomly chosen employee whose SUP score is 42.

**PART C. [33 points]** A team of Human Resource consultants, who are investigating the salaries of recent college graduates working in the area of software development, fit the following regression model to their data:

$$E(\text{sal} | \text{yr}, G, E) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 G^* E + \beta_4 \text{yr} + \beta_5 G^* \text{yr} + \beta_6 E^* \text{yr} + \beta_7 G^* E^* \text{yr},$$

where

*sal* = annual salary in tens of thousands of dollars per year;

*G* = gender (male = 0; female = 1);

*E* = educational background (computer science = 0; engineering = 1);

*yr* = completed years of work experience,

and the variables  $G^* E$ ,  $G^* \text{yr}$ ,  $E^* \text{yr}$ , and  $G^* E^* \text{yr}$  are interaction variables. The following printout shows the results of the regression analysis:

### Regression Analysis: sal versus G, E, G\*E, yr, G\*yr, E\*yr, G\*E\*yr

The regression equation is

$$\text{sal} = 60.9 + 4.51 G + 8.17 E - 1.92 G^* E + 5.44 \text{yr} - 2.09 G^* \text{yr} + 1.67 E^* \text{yr} - 0.28 G^* E^* \text{yr}$$

Predictor	Coef	SE Coef	T	P
Constant	60.909	1.948	31.27	0.000
G	4.515	2.977	1.52	0.130
E	8.174	2.964	2.76	0.006
G*E	-1.923	4.772	-0.40	0.687
yr	5.4388	0.5270	10.32	0.000
G*yr	-2.0897	0.8368	-2.50	0.013
E*yr	1.6710	0.8175	2.04	0.041
G*E*yr	-0.282	1.353	-0.21	0.835

S = 14.1002    R-Sq = 43.5%    R-Sq(adj) = 42.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	7	78280	11183	56.25	0.000
Residual Error	512	101794	199		
Total	519	180074			

1. [2 points] How many subjects are included in the data set?

2. [2 points] What is the fitted regression relationship between  $sal$  and  $yr$  for males ( $G=0$ ) who are computer science graduates ( $E=0$ )?

3. [2 points] What is the estimate of the mean starting salary for male computer science graduates, i.e., what is the estimate of  $E(sal | yr=0, G=0, E=0)$ ?

4. [3 points] Give a 95% confidence interval for the mean starting salary for males who are computer science graduates, i.e., give a 95% confidence interval for  $E(sal | yr=0, G=0, E=0)$ .

5. [2 points] What is the estimate of the rate of change in mean salary of male computer science graduates per year of work experience, i.e., how is  $E(sal | yr, G=0, E=0)$  estimated to change as  $yr$  is increased by a single unit?

6. [3 points] What is the fitted regression relationship between  $sal$  and  $yr$  for females ( $G=1$ ) who are computer science graduates ( $E=0$ )?

7. [2 points] Among computer science graduates, estimate the difference between the mean starting salary for females and the mean starting salary for males.

8. [2 points] Among computer science graduates, is there a statistically significant difference between the mean starting salary for females and the mean starting salary for males? Perform a test at the 5% level.

9. [3 points] Among computer science graduates, estimate the difference between males and females in the rate of change of mean salary per year of work experience.

10. [2 points] Among computer science graduates, is there a statistically significant difference between males and females in the rate that mean salary changes per year of work experience? Perform a test at the 5% level.

11. [3 points] Estimate the mean salary of a male engineering graduate having four years of work experience. (Justify your answer briefly.)

12. [3 points] Estimate the mean salary of a female engineering graduate having four years of work experience. (Justify your answer briefly.)

13. [4 points] What is the value of the test statistic for testing "the significance of the regression"? State the null hypothesis in this test. What is the alternative hypothesis? Would the null hypothesis be rejected in a test at the 1% level?

**PART D. [7 points]** The consultants are especially interested in testing whether the effect of gender is the same for computer science graduates as it is for engineering graduates. Consequently, they fit the "reduced" regression model

$$E(\text{sal} | \text{yr}, G, E) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 \text{yr} + \beta_4 G^* \text{yr} + \beta_5 E^* \text{yr},$$

thereby obtaining the following printout:

**Regression Analysis: sal versus G, E, yr, G\*yr, E\*yr**

The regression equation is  
 $\text{sal} = 61.3 + 3.68 G + 7.34 E + 5.47 \text{yr} - 2.21 G^* \text{yr} + 1.59 E^* \text{yr}$

Predictor	Coef	SE Coef	T	P
Constant	61.289	1.776	34.52	0.000
G	3.684	2.323	1.59	0.113
E	7.344	2.320	3.17	0.002
yr	5.4656	0.4848	11.27	0.000
G*yr	-2.2071	0.6569	-3.36	0.001
E*yr	1.5894	0.6505	2.44	0.015

S = 14.0893    R-Sq = 43.3%    R-Sq(adj) = 42.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	5	78041	15608	78.63	0.000
Residual Error	514	102033	199		
Total	519	180074			

1. [3 points] What is the value of the test statistic for testing the null hypothesis that the reduced model holds, i.e., for testing the null hypothesis that the coefficients of  $G^*E$  and  $G^*E^* \text{yr}$  are simultaneously zero in the full model of PART A? (Justify your answer briefly.)

2. [4 points] Would the reduced model be rejected in favor of the full model of PART A? Perform a test at the 5% level. (Justify your answer briefly.)