CONCORDIA UNIVERSITY
Department of Economics

ECON 222/2 SECTIONS A and AA
STATISTICAL METHODS II
FALL 2014 – ASSIGNMENT 1
Due: Monday, September 29, before 4:00 pm

1. **(15 marks)** Download the Excel file A1Q1.xlsx from Moodle. The spreadsheet contains 30 randomly generated numbers.

   a. Use the descriptive statistics feature in Excel to calculate the descriptive statistic. Attach the Excel output. (**3 marks**)

Mean	*Variance*	*StdDev*	*Maximum*	*Minimum*
0.543806765	*28.48439998*	*5.428316559*	*11.05244576*	*-9.102450999*

   b. Construct a 95-percent confidence interval for the population variance and the population mean. (**4 marks**)

   For the population variance:

$$\frac{(n-1)S^2}{\chi^2_{0.025}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{0.975}}$$

$$S^2 = 28.4843,\ n-1 = 29,\ \chi^2_{0.975} = 13.844,\ \chi^2_{0.025} = 41.923$$

   Therefore,

$$19.704 \leq \sigma^2 \leq 59.67$$

   For the population mean:

$$\bar{X} - t_{0.025,29} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{0.025,29} \cdot \frac{S}{\sqrt{n}}$$

$$\bar{X} = 0.5438,\ n = 30,\ t_{0.025,29} = 2.05,\ S = 5.428$$

$$0.5438 - 2.05 \cdot \frac{5.428}{5.47} \leq \mu \leq 0.5438 + 2.05 \cdot \frac{5.428}{5.47}$$

   Therefore,

$$-1.49 \leq \mu \leq 2.578$$

c. At the 90-percent level of confidence, test whether the population variance equals 25. Carefully state the null and alternative hypotheses, calculate the appropriate test statistic, state the appropriate critical value and explain your decision. (**4 marks**)

$H_0 : \sigma^2 = 25$, $H_1 : \sigma^2 \neq 25$

The test statistic is $\chi^2 = (n-1)\frac{S^2}{\sigma^2} = 29 \cdot \frac{28.4843}{625} = 1.32$. *We reject the null hypothesis if* $\chi^2_{1-\frac{\alpha}{2},n-1} \leq \chi^2 \leq \chi^2_{\frac{\alpha}{2},n-1}$. *From the tables, we know* $\chi_{0.975,29} = 16.047$ *and* $\chi_{0.025,29} = 45.722$.

Since the test statistic does not lie between the critical values, we cannot reject the null hypothesis.

d. At the 99-percent level of confidence, test whether the population mean equals zero. Carefully state the null and alternative hypotheses, calculate the appropriate test statistic, state the appropriate critical value and explain your decision. (**4 marks**)

$H_0 : \mu = 0$, $H_1 : \mu \neq 0$

The test statistic is $t = \frac{X - \mu^2}{S/\sqrt{n}} = \frac{0.5438 - 0}{5.428/\sqrt{30}} = 0.5487$. *We reject the null hypothesis if* $|t| > t_{\frac{\alpha}{2},n-1}$. *From the tables, we know* $t_{0.005,29} = 2.756$.

Since the test statistic does not lie beyond the critical value, we cannot reject the null hypothesis.

2. **(2 marks each = 22 marks)** Write the following summations in expanded form.

a. $\sum_{k=1}^{5} 3k = 3\sum_{k=1}^{5} k = 3\sum_{k=1}^{5} k = 3\cdot(1+2+3+4+5) = 45$

b. $\frac{1}{4}\sum_{m=1}^{4} x_m = \frac{1}{4}(x_1 + x_2 + x_3 + x_4)$

c. $\sum_{j=4}^{n} \frac{j}{j+1} = \frac{4}{4+1} + \frac{5}{5+1} + \ldots + \frac{n}{n+1}$

d. $\sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \frac{(O_3 - E_3)^2}{E_3} + \frac{(O_4 - E_4)^2}{E_4}$

e. $\sum_{j=3}^{7} 3^{j-1} = 3^{3-1} + 3^{4-1} + 3^{5-1} + 3^{6-1} + 3^{7-1} = 9 + 27 + 81 + 243 + 729 = 1089$

f. $\frac{1}{5}\sum_{k=1}^{5} x_k = \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5)$

g. $2\sum_{t=2}^{m} \frac{t-1}{t+1} = 2\cdot\left(\frac{2-1}{2+1} + \frac{3-1}{3+1} + \ldots + \frac{m-1}{m+1}\right)$

h. $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + \bar{x}^2\sum_{i=1}^{n} 1\right)$
$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)$

i. $\sum_{l=1}^{5} 2l^2 = 2\cdot(1^2 + 2^2 + 3^2 + 4^2 + 5^2) = 110$

j. $\sum_{k=1}^{3} 5^{k+1} = 5^{1+1} + 5^{2+1} + 5^{3+1} = 25 + 125 + 625 = 775$

k. $\sum_{i=1}^{k+1} \frac{3(i-2)}{i+2} = 3\sum_{i=1}^{k+1} \frac{i-2}{i+2} = 3\cdot\left(\frac{1-2}{1+2} + \frac{2-2}{2+2} + \ldots + \frac{k-1}{k+3}\right)$

3. **(26 marks)** Use the provided data to answer the following questions.

X_i	Y_i	X_i^2	X_iY_i	$\hat{Y}_i$	e_i	e_i^2	$Y_i - \bar{Y}_i$	$(Y_i - \bar{Y}_i)^2$	$\hat{Y}_i - \bar{Y}_i$	$(\hat{Y}_i - \bar{Y}_i)^2$
7	2	49	14	1.64	0.36	0.13	-2.43	5.90	-2.78	7.75
4	4	16	16	3.98	0.02	0.00	-0.43	0.18	-0.45	0.20
6	2	36	12	2.42	-0.42	0.18	-2.43	5.90	-2.00	4.02
2	5	4	10	5.54	-0.54	0.29	0.57	0.33	1.11	1.24
1	7	1	7	6.32	0.68	0.46	2.57	6.61	1.89	3.59
1	6	1	6	6.32	-0.32	0.10	1.57	2.47	1.89	3.59
3	5	9	15	4.76	0.24	0.06	0.57	0.33	0.33	0.11
$\sum_{i=1}^{7} X_i =$ 24	$\sum_{i=1}^{7} Y_i =$ 31	$\sum_{i=1}^{7} X_i^2 =$ 116	$\sum_{i=1}^{7} X_iY_i =$ 80			$\sum_{i=1}^{7} e_i^2 =$ 1.22		$\sum_{i=1}^{7} (Y_i - \bar{Y}_i)^2 =$ 21.71		$\sum_{i=1}^{7} (\hat{Y}_i - \bar{Y}_i)^2 =$ 20.49

a. Fill in the above table. (**2 marks for the $\hat{Y}_i$ column; 1 mark for each other column.**)

b. Calculate $\bar{X}$. (**1 mark**)

$$\bar{X} = \frac{1}{7}\sum_{i=1}^{7} X_i = \frac{24}{7}$$

c. Calculate $\bar{Y}$. (**1 mark**)

$$\bar{Y} = \frac{1}{7}\sum_{i=1}^{7} Y_i = \frac{31}{7}$$

i. Calculate *ESS*. (**1 mark**)

$$ESS = \sum_{i=1}^{7} (\hat{Y}_i - \bar{Y})^2 = 20.49$$

j. Calculate *TSS*. (**1 mark**)

$$TSS = \sum_{i=1}^{7} (Y_i - \bar{Y})^2 = 21.71$$

d. Calculate $\text{cov}(X,Y)$. (**2 marks**)

$$\text{cov}(X,Y) = \sum_{i=1}^{7} X_i Y_i - n\overline{X}\overline{Y} = 80 - 7 \cdot \frac{24}{7} \cdot \frac{31}{7} = -\frac{184}{7}$$

e. Calculate $\text{var}(X)$. (**2 marks**)

$$\text{var}(X) = \sum_{i=1}^{7} X_i^2 - n\overline{X}^2 = 116 - 7 \cdot \left(\frac{24}{7}\right)^2 = \frac{236}{7}$$

f. Calculate $\hat{\beta}_1$. (**1 mark**)

$$\hat{\beta}_1 = \frac{\text{cov}(X,Y)}{\text{var}(X)} = \frac{-184/7}{236/7} = -\frac{46}{59}$$

g. Calculate $\hat{\beta}_0$. (**1 mark**)

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = \frac{31}{7} - \left(-\frac{46}{59}\right) \cdot \frac{24}{7} = \frac{2933}{413}$$

h. Calculate *RSS*. (**1 mark**)

$$RSS = \sum_{i=1}^{7} e_i^2 = 1.22$$

k. Calculate R^2. (**1 mark**)

$$R^2 = \frac{ESS}{TSS} = \frac{20.49}{21.71} = 0.944$$

l. Calculate *r*. (**1 mark**)

$$r = -\sqrt{R^2} = -\sqrt{0.944} = -0.971$$

m. Calculate *MSE*. (**1 mark**)

$$MSE = \frac{1}{n-k-1} \sum \left(Y_i - \hat{Y}_i\right)^2 = \frac{RSS}{n-2} = \frac{1.22}{5} = 0.244$$

n. Calculate *MSR*. (**1 mark**)

$$MSR = \frac{1}{k} \sum \left(\hat{Y}_i - \overline{Y}\right)^2 = \frac{ESS}{1} = 20.49$$

o. Calculate *s*. (**1 mark**)

$$s = \sqrt{\frac{\sum (Y_i - \hat{Y}_i)^2}{n-k-1}} = \sqrt{MSE} = \sqrt{0.244} = 0.494$$

4. **(23 marks)** A marketing manager wants to establish the relationship between the number of cereal boxes sold, *B*, and the shelf space in square feet devoted to them, *S*. This relationship is assumed to be linear of the form shown below:

$$B_i = \beta_0 + \beta_1 S_i + \varepsilon_i, \text{ for } i = 1,\ldots,n$$

a. Predict and justify the sign that you would expect for β_1. (**2 marks**)

The sign of β_1 should be positive. Increasing the shelf space that is devoted to cereal boxes should lead to more sales, as an expanded space enhances the visibility and therefore the appeal of the product to the customers.

b. Download the Excel file A1Q4.xlsx from Moodle. The spreadsheet contains weekly sales of cereal boxes and the shelf space in square feet devoted to them for 14 urban grocery stores. Estimate the regression in Excel, attach your Excel output and write down the estimated regression equation. (**4 marks**)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.746
R Square	0.557
Adjusted R Square	0.520
Standard Error	56.935
Observations	14

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	48960.19	48960.19	15.10	0.002163
Residual	12	38899.81	3241.65		
Total	13	87860.00			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	62.25	36.22	1.72	0.11	-16.67	141.17
Shelf Space	15.97	4.11	3.89	0.00	7.02	24.92

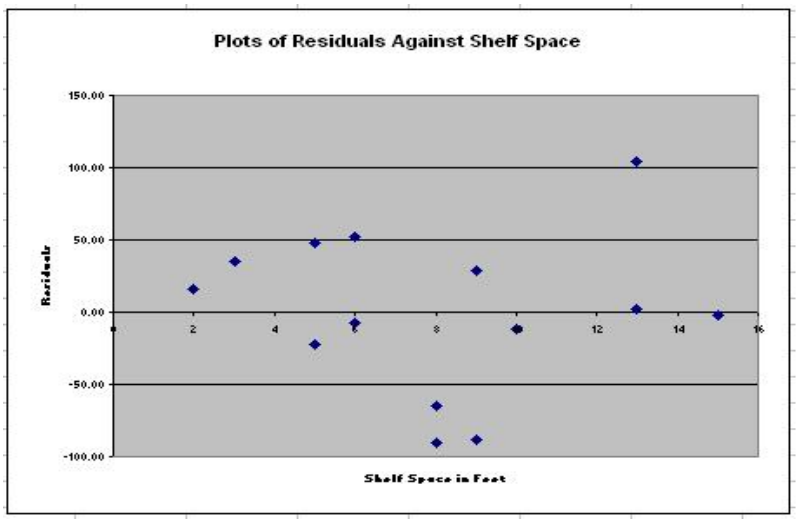
The estimated equation is $\hat{B}_i = 62.25 + 15.97 S_i$.

c. Interpret the estimated coefficients. (**2 marks**)

The estimated coefficient β_0 gives the estimated number of cereal boxes that would be sold if zero shelf space were devoted to it (ie, people ask for this brand because of its reputation).

The estimated coefficient β_1 says that a 1 square-foot increase in shelf space devoted to cereal boxes should result in an increase in 16 cereal boxes sold.

d. Calculate the residuals and plot them against the independent variable. Attach your Excel graph. (**4 marks**)



e. Delete the 8th, 11th, 12th and 13th observations and re-estimate the regression in Excel. Attach your Excel output and write down the new estimated regression equation. (**4 marks**)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.924
R Square	0.854
Adjusted R Square	0.836
Standard Error	25.760
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	31001.47	31001.47	46.72	0.000133
Residual	8	5308.63	663.58		
Total	9	36310.10			

	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	92.06	17.03	5.40	0.00	52.78	131.34
Shelf Space	13.82	2.02	6.84	0.00	9.16	18.48

The estimated equation is $\hat{B}_i = 92.06 + 13.82 S_i$

f. Graph the estimated regression equations lines from (b) and (e). Do they intersect and, if so, at what point? (**3 marks**)

The two estimated equations are $\hat{B}_i = 62.25 + 15.97S_i$ *are* $\hat{B}_i = 92.06 + 13.82S_i$.

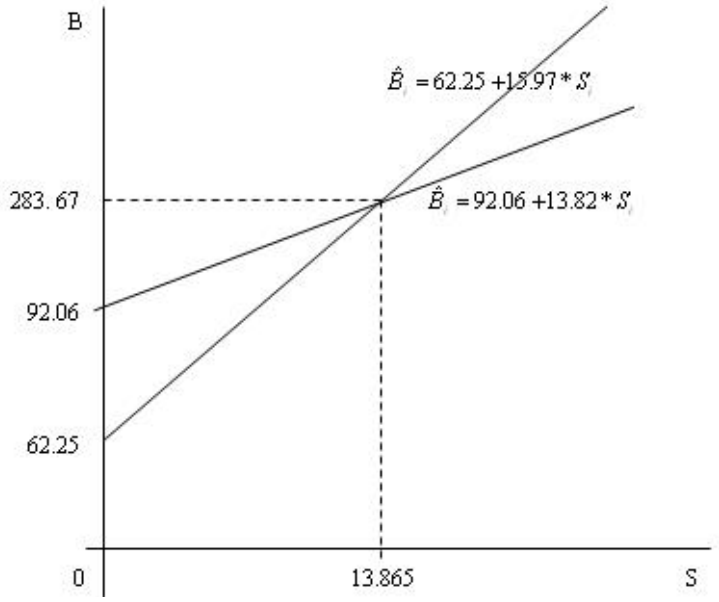
Set the two equations equal and solve for S:

$$62.25 + 15.97S_i = 92.06 + 13.82S_i \Rightarrow 2.15S_i = 27.81 \Rightarrow S_i = \frac{29.81}{2.15} = 13.87$$

Substitute the value of S into either equation and solve for B:

$$\hat{B}_i = 62.25 + 15.97 \cdot 13.87 = 92.06 + 13.82 \cdot 13.87 = 283.67$$

Intersection of the two estimated regression lines



g. Compare the estimated coefficients in (b) and (e). Briefly explain why they differ? (**2 marks**)

The cofficients differ in equations in (b) and (e) because they are derived from different samples.

h. Compare the R^2 coefficients in (b) and (e). Which one is better and why? (**2 marks**)

The R^2 in the second model (0.854) is better than the first R^2 (0.557) because the second model removed some outlier observations that were adding large residuals.

5. **(14 marks)** A farmer wants to establish the relationship between the per-acre yield of corn, *Y* (measured in bushels per acre), and the average July temperature in Manitoba, *X* (measured in degrees Fahrenheit), using information from the past eight years. Preliminary analysis of the sample data produces the following information:

$$\sum Y_i = 848 \qquad \sum X_i = 736 \qquad \sum Y_i^2 = 91334 \qquad \sum X_i^2 = 67852$$

$$\sum X_i Y_i = 78429 \qquad \sum e_i^2 = 227.65 \qquad \sum (X_i - \overline{X})(Y_i - \overline{Y}) = 413$$

Use the above sample information to answer all the following questions. Show explicitly all formulas and calculations.

a. Calculate the estimated intercept coefficient, β_0, and slope coefficient, β_1. (**4 marks**)

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{\sum X_i^2 - n\overline{X}^2} = \frac{78429 - 8 \cdot \frac{736}{8} \cdot \frac{848}{8}}{67852 - 8 \cdot \left(\frac{736}{8}\right)^2} = \frac{78429 - 78016}{67852 - 67712} = \frac{413}{140} = 2.95$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = \frac{848}{8} - 2.95 \cdot \frac{736}{8} = 106 - 2.95 \cdot 92 = -165.4$$

b. Write down the estimated regression equation and interpret the value of the slope coefficient. (**4 marks**)

The estimated value of β_1 in the estimated equation, $\hat{Y}_i = -165.4 + 2.95X_i$, says that an increase in temperature by one degree Fahrenheit should result in an increase of 2.95 bushels of corn per acre.

c. Calculate the estimated error variance, $\hat{\sigma}^2$. (**2 marks**)

$$S_e^2 = \frac{RSS}{n-k-1} = \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{n-k-1} = \frac{\sum e_i^2}{n-k-1} = \frac{227.65}{8-1-1} = 37.9417$$

d. Calculate the coefficient of determination, R^2. Briefly explain what it means. (**4 marks**)

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum \left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\sum e_i^2}{\sum Y_i^2 - n\bar{Y}^2} = 1 - \frac{227.65}{91334 - 8 \cdot \left(\frac{848}{8}\right)^2} = 1 - \frac{227.65}{91334 - 89888} = 0.843$$

The coefficient of determination means that our model's formulation can explain 84.3 percent of the variation in corn yields per acre.