CONCORDIA UNIVERSITY
Department of Economics
ECON 222 *Statistical Methods II*
Instructor: Mesbah F. Sharaf
Winter 2012-2013
Key answers for Assignment 1

1)
a) See page 54
b) See section 1.3
c) See section 3.1
d) See section 2.1

2) See sections 1.2 and 1.3
3) See section 2.4

Question 4

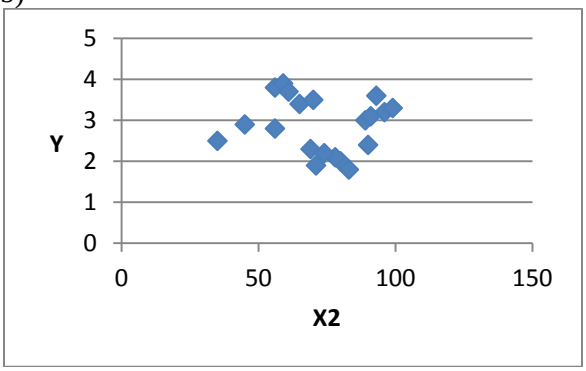
a)

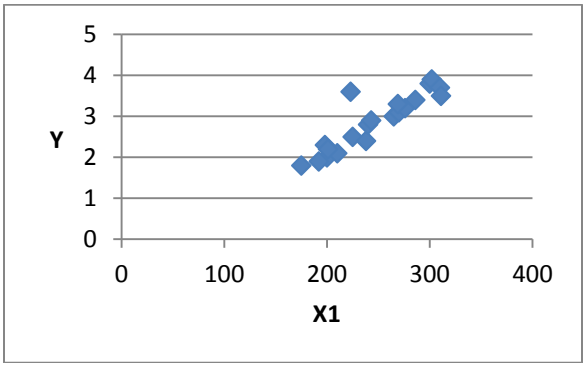
Mean	2.87
Standard Error	0.151848262
Median	2.95
Mode	#N/A
Standard Deviation	0.679086073
Sample Variance	0.461157895
Kurtosis	-1.36087601
Skewness	-0.09465304
Range	2.1
Minimum	1.8
Maximum	3.9
Sum	57.4
Count	20

According to the findings in (a) above,

- The distribution of the observations in the sample is not symmetrical
- The kurtosis indicator of our distribution is -1.36 which implies that our distribution has a shorter peak as compared to the standard normal distribution
- The skewness indicator of our distribution is -0.094, which implies that, compared to the standard normal distribution, our distribution is negatively skewed (skewed to the left)

b)





C)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.900837
R Square	0.811508
Adjusted R Square	0.801036
Standard Error	0.302909
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	7.110429	7.110429	77.49454	6.11E-08
Residual	18	1.651571	0.091754		
Total	19	8.762			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.64095	0.404542	-1.58439	0.130515	-1.49087	0.208958	-1.49087	0.208958
X 1	0.014232	0.001617	8.803098	6.11E-08	0.010835	0.017628	0.010835	0.017628

D)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.076606
R Square	0.005869
Adjusted R Square	-0.04936
Standard Error	0.695644
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	1	0.05142	0.05142	0.106257	0.748206
Residual	18	8.71058	0.483921		
Total	19	8.762			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	3.085727	0.679834	4.538942	0.000254	1.657449	4.514005	1.657449	4.514005
X 2	-0.00296	0.009066	-0.32597	0.748206	-0.022	0.016091	-0.022	0.016091

E)

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.901045
R Square	0.811883
Adjusted R Square	0.789751
Standard Error	0.311381
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	2	7.113716	3.556858	36.68457	6.8E-07
Residual	17	1.648284	0.096958		
Total	19	8.762			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-0.70389	0.538298	-1.30762	0.208411	-1.8396	0.43182	-1.8396	0.43182
X 1	0.014264	0.001671	8.534566	1.5E-07	0.010738	0.017791	0.010738	0.017791
X 2	0.000751	0.004081	0.184129	0.856091	-0.00786	0.009362	-0.00786	0.009362

f) Beta1: spending an extra hour studying on the internet will increase the GPA by 0.014
Beta 2: as the student increase his attended classes every week by 1 percent, GPA will increase by 0.000751.
R^2 : 81.18 % of the variation of Y (GPA) around its mean is explained by the regression model.

g) Model 2 is out of consideration since beta 2 has the wrong sign and is not statistically significant (very low t statistic) and adjusted R^2 is very small. Model 1 is better than model 3 since it has higher R^2 adjusted, the sign of the coefficients are consistent with the theory (all coefficients are positive) and x2 is not statistically significant in model 3. Moreover, model 2 has a higher F statistic than model 3.

Question 5

A realtor wants to establish the relationship between the number of weeks homes are on the market prior to sale, (*Y*), and the difference of the asking price from the municipal taxable value in thousands, (*X*).

$\sum Y_i = 168$ $\sum X_i = 1624$ $\sum Y_i^2 = 2436.02$ $\sum X_i^2 = 211856$

$\sum X_i Y_i = 22131.7$ $\sum e_i^2 = 122.255$ n = 14

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

a. Compute OLS estimates of the intercept coefficient β_0 and the slope coefficient β_1.

We plug the figures in the formulas for the estimated coefficients and get

$$\hat{\beta}_1 = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{\sum X_i^2 - n\overline{X}^2} = \frac{22131.7 - 14 * \frac{168}{14} * \frac{1624}{14}}{211856 - 14 * \left(\frac{1624}{14}\right)^2} = \frac{22131.7 - 19488}{211856 - 188384} = \frac{2643.7}{23472} = 0.11263$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = \frac{168}{14} - 0.11263 * \frac{1624}{14} = 12 - 0.11263 * 116 = -1.0651$$

$$\hat{Y}_i = -1.0651 + 0.11263 * X_i$$

b. Interpret the slope coefficient estimate you calculated in part (a) -- i.e., explain in words what the numeric value you calculated for β_1 means.

The estimated value of the β_1 coefficient says that for every additional thousand dollar above the municipal valuation asked, the sale of a house is delayed by 0.113 weeks, or, for every 8.9 additional thousand dollars above the municipal valuation asked, the sale of a house is delayed by one week, on average.

c. Calculate an estimate of σ^2, the error variance.

We plug the figures in the formula for the variance of the residuals and get

$$S_e^2 = \frac{RSS}{n-k-1} = \frac{\sum \left(Y_i - \hat{Y}_i\right)^2}{n-k-1} = \frac{\sum e_i^2}{n-k-1} = \frac{122.255}{14-1-1} = 10.188$$

d. Calculate an estimate of the variance and standard error of the slope coefficient β_1.

We plug the figures in the formula for the variance of β_1 and get

$$SE_{\hat{\beta}_1}^2 = \frac{S_e^2}{\sum (X_i - \overline{X})^2} = \frac{S_e^2}{\sum X_i^2 - n\overline{X}^2} = \frac{10.188}{211856 - 14 * \left(\frac{1624}{14}\right)^2} = \frac{10.188}{211856 - 188384} = 0.00043405$$

And *the standard error of β_1 is the squared root of the variance:*

$$SE_{\hat{\beta}_1} = \sqrt{0.00043405} = 0.020834$$

e. Compute the value of the coefficient of determination for the estimated OLS sample regression equation R^2. Briefly explain what the calculated value of R^2 means.

We plug the figures in the formula for the Coefficient of Determination R^2 and get

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^{\,2}}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{\sum e_i^2}{\sum Y_i^2 - n\bar{Y}^2}$$

$$\textit{Therefore, } R^2 = 1 - \frac{122.255}{2436.02 - 14 * \left(\frac{168}{14}\right)^2} = 1 - \frac{122.255}{2436.02 - 2016} = 1 - \frac{122.255}{420.02} = 0.70893$$

The calculated value of the Coefficient of Determination gives us a measure of the goodness of fit of our data. Our model's formulation enables us to explain 70.9% of the variations observed in delays in house sales (in terms of weeks).

f. Test the null hypothesis H_0: *$\beta_1 = 0$* against an alternative hypothesis H_A of your choice.

It is expected that the relationship between the number of weeks homes are on the market prior to sale and the difference of the asking price from the municipal taxable value *to be positive. This means that the higher the difference in the asking price from the municipal valuation the longer it will take to be sold. Therefore, this is a positive one-sided test that will be run at a 5% level of significance.*

H_0: $\beta_k \leq 0$ (the values we do not expect to be true)
H_A: $\beta_k > 0$ (the values we expect to be true)

The critical t for 12 degrees of freedom at 5% level of significance is 1.782

The decision rule is: Reject H_0 if | t-estimated| > |t-critical = 1.782|

$$t_{n-k-1} = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}} \text{ or } \qquad t_{12,0.05} = \frac{0.11263 - 0}{\sqrt{0.00043405}} = \frac{0.11263}{0.020834} = 5.4061 > 1.782$$

The t-estimated is larger than the critical t-ratio and the coefficient is positive. Therefore, we reject the null hypothesis and conclude that the estimated coefficient is significantly different than zero. This implies that the independent variable to which this coefficient is attached (difference of asking price from municipal valuation) is significant in explaining differentials in delays in home sales.

g. Test whether you could accept the hypothesis that the true slope coefficient could be 0.15 or more.

This is a one-sided and the hypotheses to be tested are set as follows:

H_0: $\beta_1 \geq 0.15$
H_A: $\beta_1 < 0.15$

The critical value of t is -1.782, *and the test is set as* $t_{est} = \dfrac{\hat{\beta}_1 - 0.15}{s_{\hat{\beta}1}}$

Rule: Reject H_0 *if* $t_{est} < -1.782$

$$t_{est} = \frac{0.11263 - 0.15}{0.020834} = \frac{-0.03737}{0.020834} = -1.7937 < -1.782$$

The estimated t value is smaller than the critical t. Therefore we reject the null hypothesis and we conclude that the true slope coefficient could not be 0.15 or higher, at the 5% level of significance (with 95% probability).

We could "accept" the null hypothesis at the 4.9% level of significance.

h. Make an interval estimate of the true β_1 coefficient.

We plug the figures in the formula for a 95% confidence interval for the true coefficient β_1 *and get*

$$\hat{\beta} - t_{n-k-1} SE_{\hat{\beta}} \leq \beta \leq \hat{\beta} + t_{n-k-1} SE_{\hat{\beta}}$$

$0.11263 - 2.179*0.020834 \leq \beta_1 \leq 0.11263 + 2.179*0.020834$ or $0.06723 \leq \beta_1 \leq 0.15803$

We are 95% confident that the true marginal effect of the difference in asking price from the municipal valuation on sale delays is found between 0.067 and 0.158 weeks per thousands of dollars difference. Or, for every additional $10 thousand difference in asking price from the municipal valuation the sale can be delayed from about 5 to 11 days.

6)

a) All the coefficients have the expected signs.

b) No. The sizes of the estimated coefficients depend on the units at which their respective variables are expressed. Although the size of the coefficient of (NO) is bigger than that of (TX) , we cannot say that (NO) has a bigger effect. Given that the two variables are measured in different units, we cannot directly compare their influence based on the size of their coefficients. Similarly, the statistical significance (t-ratios) or relative strength of the coefficients cannot be used to compare contributions to the explanatory power of the model.

c)

A: P=9.561-0.078*65-0.913*8+5.1*6-1.012*2-0.07*25-0.556*30=7.33

B: P=9.561-0.078*8-0.913*5+5.1*6-1.012*10-0.07*60-0.556*8= 16.2

D) **Develop and test appropriate hypotheses about the individual slope coefficients at the 5% significance level.**

- *For the first, second, fourth, fifth and sixth coefficients whose impact is expected to be unambiguously negative, the test will be one-sided and the hypotheses to be tested are set as follows:*

  H_0: $\beta_k \geq 0$
  H_A: $\beta_k < 0$

The critical value of t for 499 degrees of freedom is 1.645, and the test is set as

$t_{est} = \dfrac{\hat{\beta}_k - 0}{s_{\hat{\beta}_k}}$ *Rule: Reject* H_0 *if* $|t_{cal}| > t_{critical}$ & t_{cal} *has the same sign implied by* H_1

a. For the coefficient of CR $t_{est} = \dfrac{0.078 - 0}{0.052} = 1.5 < 1.645$

The absolute value of the calculated t is smaller than the critical t , then we "accept" the null hypothesis at 5% and we conclude that the estimated coefficient does not have impact on the housing prices at the significance level of 5 %.

b. For the coefficient of NO $t_{est} = \dfrac{0.913 - 0}{0.378} = 2.4153 > 1.645$

The absolute value of the calculated t is bigger than the critical t and the coefficient has the right sign (negative as expected). Therefore, we cannot accept the null hypothesis and we conclude that the estimated coefficient is strong. This implies that the level of nitrous oxide in the air relates negatively to the median housing prices.

c. For the coefficient of DS $t_{est} = \dfrac{1.012 - 0}{0.179} = 5.654 > 1.645$

The absolute value of the calculated t is bigger than the critical t and the coefficient has the right sign (negative as expected). Therefore, we cannot accept the null hypothesis and we conclude that the estimated coefficient is strong. This implies that weighted average distance from employment centers is negatively related to the median housing prices.

d. For the coefficient of TX $t_{est} = \dfrac{0.070 - 0}{0.021} = 3.33 > 1.684$

The absolute value of the calculated t is bigger than the critical t and the coefficient has the right sign (negative as expected). Therefore, we cannot accept the null hypothesis and we conclude that the estimated coefficient is strong. This implies property tax is negatively related to the median housing prices.

e. For the coefficient of LS $t_{est} = \dfrac{0.556 - 0}{0.051} = 10.9 > 1.684$

The absolute value of the calculated t is bigger than the critical t and the coefficient has the right sign (negative as expected). Therefore, we cannot accept the null hypothesis and we conclude that the estimated coefficient is strong. This implies that percentage of people of lower status, suggests lower buying power of the community which relates negatively to the median housing prices.

- *For the coefficient of average number of rooms in houses is expected to be unambiguously positive, the test will be one-sided and the hypotheses to be tested are set as follows:*
  H_0: $\beta_k \leq 0$ (the values we do not expect to be true)
  H_A: $\beta_k > 0$ (the values we expect to be true)

  The critical value of t is 1.645 and the test is set as $t_{est} = \dfrac{\hat{\beta}_k - 0}{s_{\hat{\beta}_k}}$

  Rule: Reject H_0 if $t_{cal} > 1.645$

  For the coefficient of RM $t_{est} = \dfrac{5.1 - 0}{0.432} = 11.81 > 1.645$

  The absolute value of the calculated t is bigger than the critical t and the coefficient has the right sign (positive as expected). Therefore, we cannot accept the null hypothesis and we conclude that the estimated coefficient is strong. This implies that the average number of rooms of houses contribute positively towards the median housing prices.

e)

This is a test about the validity of the whole model and consists of a simultaneous test of significance for all slope coefficients taken together. For this test we use the F-ratio.

H_0: $\beta_{CR} = \beta_{NO} = \beta_{RM} = \beta_{DS} = \beta_{TX} = \beta_{LS} = 0$
H_A: $\beta_{CR} \neq \beta_{NO} \neq \beta_{RM} \neq \beta_{DS} \neq \beta_{TX} \neq \beta_{LS} \neq 0$

Reject H_0 if $F_{cal} > F_{6, 499, 0.05} = 2.10$

$$F_{est} = \frac{ESS/k}{RSS/(n-k-1)} = \frac{R^2}{1-R^2} * \frac{n-k-1}{k} = \frac{0.667}{0.33} * \frac{506-6-1}{6} = 168.09 > 2.10$$

Therefore, we reject the null hypothesis and we conclude that the model at hand is valid.

f) $S_Y{}^2 = \dfrac{\sum\left(Y_i - \bar{Y}\right)^2}{n-1}$ where $\sum\left(Y_i - \bar{Y}\right)^2 = TSS = ESS + RSS$

Knowing that RSS = 14259.5 as well as that RSS represents (1- R^2)% of TSS,

We compute TSS = 14259.5 / 0.33 = 42821.32

Therefore, $S_Y{}^2 = \dfrac{\sum\left(Y_i - \bar{Y}\right)^2}{n-1} = \dfrac{42821.32}{506-1} = 84.79$ *and* $S_Y = \sqrt{84.79} = 9.208$

g) Holding of the Classical Assumptions means that we can trust our estimates being BLUE, as coming out of best linear unbiased estimators. This means that we are confident that our estimates are unbiased and very close to the true population parameters. However, this does not necessarily imply that the estimated coefficients are the true population parameters, which we can never know. Thus, the estimated marginal effect of 5.1 is the best estimate of the true value of the coefficient of the number of rooms that we could have come up with.

H)

R^2 is but one of the indicators of the strength of the regression results. Specifically, R^2 gives us an idea of the extent of the sample variance of the dependent variable that we have succeeded in explaining through the use of our model. And although it appears here that only a small part of this variation has been explained through our model, we have, however, established that the estimated coefficients are significantly different than zero. In addition, the validity of the overall model has been established with the F-test. Therefore, the variables that are included in the model have already been validated as offering good explanation (reasons) for the variation observed in the housing prices.

In practice, a high R^2 (75%-95%) is expected in the estimation of economic or business models using time series in order for them to be considered acceptable. In the case of social research (as is our case), however, R^2 of 50% is a rarity, while the typical models carry a much smaller R^2 number. This, however, does not necessarily imply that we cannot use those estimated models to measure relationships or make predictions.