CSC591: Foundations of Data Science
MidTerm-1.

Date: 11/05/15; Total Exam Time: 1 H 05 Min.
Start Time: 12.55pm; End Time: 2.00pm

Answer Key

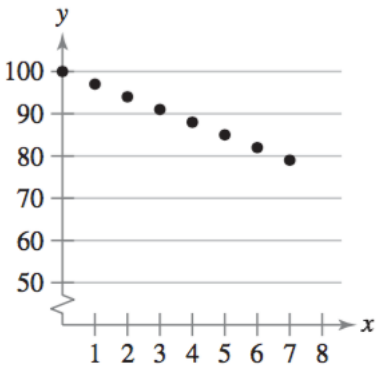
Notes

- Answer in the space provided
- You can use additional blank paper for spill over solution(s).
- Bonus question accounts for 4%.
- Read paper first, and start answering easy (and less time consuming) questions first.
- Attempt all questions. Partial grade will be awarded for incomplete answers, but you will get 0 if you don't attempt.
- Don't look at your neighbor's answers (if caught, both will be asked to leave exam hall instantly, may lead to other penalties as well).
- Don't ask questions after exam starts, if you are in doubt, simply make your best judgment (write it down clearly), in such cases, instructor/TA's decision is final.
- Only allowed items:
    - 1-page of written/typed notes (standard US A4 size paper; you can use both sides) – Please submit this page along with answer book.
    - Regular calculator
- Follow any oral instructions provided by the TA during examination.
- Total pages = 14.

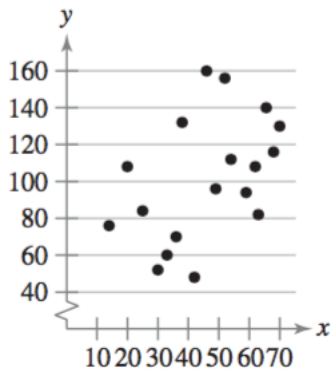
Q#	Max Points	Your Score
1	40	
2	25	
3	35	
4	30 points (3% grade)	

1. Regression (40 points)

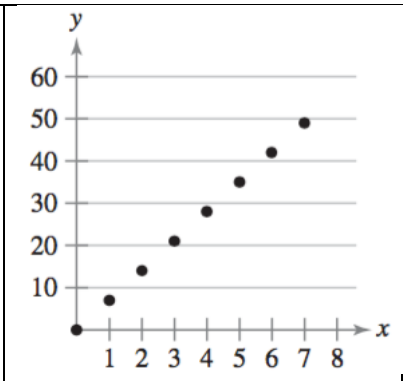
(a) Assign each plot with one of the unique labels: (i) strong positive correlation, (ii) strong negative correlation, (iii) perfect positive correlation, (iv) perfect negative correlation, (v) weak positive correlation, and (vi) no correlation. (6 points)



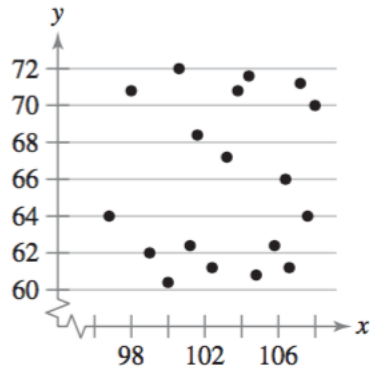
(iv)



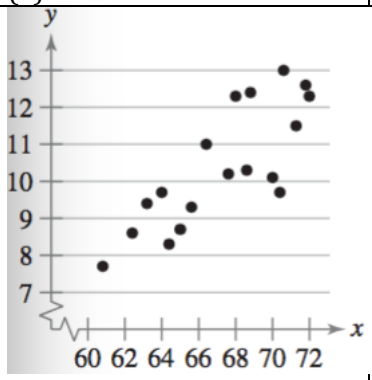
(v)



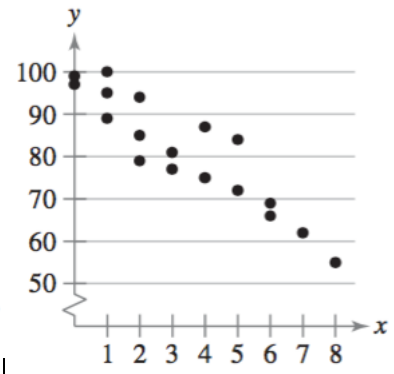
(iii)



(vi)



(i)



(ii)

(b) Which assumption of linear regression is violated for time series data (e.g., financial data (stocks), meteorological data (e.g., temperature)). (2 points)

Independence (lack of autocorrelation) of errors

(c) For the given data:

x_i	y_i	x_i^2	$x_i * y_i$	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1.6	428.2	2.56	685.12	416.1322	12.0678	145.631797
3.6	828.8	12.96	2983.68	808.4362	20.3638	414.68435
4.9	1214.2	24.01	5949.58	1063.4338	150.7662	22,730.45
1.1	444.6	1.21	489.06	318.0562	126.5438	16,013.33
0.9	264	0.81	237.6	278.8258	-14.8258	219.804346
2.9	415.3	8.41	1204.37	671.1298	-255.8298	65,448.89
2.7	571.8	7.29	1543.86	631.8994	-60.0994	3611.93788
2.3	454.9	5.29	1046.27	553.4386	-98.5386	9709.85569
1.6	358.7	2.56	573.92	416.1322	-57.4322	3298.4576
1.5	573.5	2.25	860.25	396.517	176.983	31,322.98

(1) Compute the regression equation (15 points)

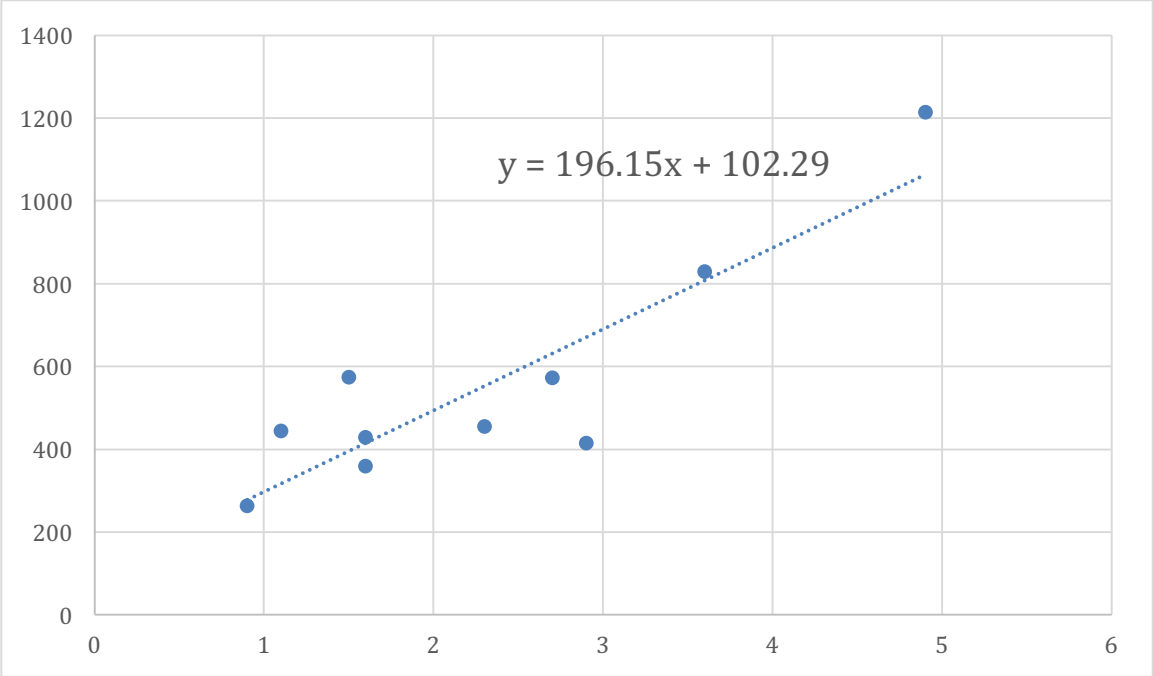
$\Sigma x = 23.1$
$\Sigma y = 5554$
$\Sigma x^2 = 67.35$
$\Sigma xy = 15573.71$

$a = ((\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)) / (n(\Sigma x^2) - (\Sigma x)^2)$
$= (5554 * 67.35 - 23.1 * 15573.71) / (10 * 67.35 - 23.1^2)$
$= 102.29$

$b = (n(\Sigma xy) - (\Sigma \mathbf{x})(\Sigma y)) / (n(\Sigma x^2) - (\Sigma x)^2)$
$= (10 * 15573.71 - 23.1 * 5554) / (10 * 67.35 - 23.1^2)$
$= 196.15$

$y'(x) = a + bx = 102.29 + 196.15x$

(2) Draw scatter plot and show regression equation (5 points)



(3) Find the coefficient of determination (5 points)

$\overline{y} = 555.4$

$R^2 = \Sigma(y' - \overline{y})^2 / \Sigma(y - \overline{y})^2 = 538{,}235.27 / 691{,}151.16 = 0.779$

(4) What is the interpretation of the answer given in (3) (2 points)

The coefficient of determination gives the fraction of the variance in the data explained by the linear fit. Specifically, because the value is close to 1, the regression line gives a good fit in this case.

(5) Find the standard error of estimate (5 points)

$s_{est} = \text{sqrt}(\Sigma(y - y')^2 / (n - 2)) = \text{sqrt}(152{,}916 / 8) = 138.26$

2. Feature Selection/Dimensionality Reduction (25 points)

(a) Outline algorithm to rank attributes using information theory measures (as discussed in the class) (5 points)
For each attribute:
- Split the data into subsets T based on the value of that attribute
- Calculate the entropy of each subset, H(T)
- Average the entropy of the subsets, weighted by the fraction of items in each set: $\Sigma_i P_i H(T_i)$
Lower average entropies indicate better (more informative) attributes

(b) Given the following two dimensional data, find the first principal component (20 points)

x1 = 1, 0, -1
x2 = -1, 1, 0

$\text{Var}(x1) = (1^2 + (-1)^2)/3 = 2/3$
$\text{Var}(x2) = ((-1)^2 + 1^2)/3 = 2/3$
$\text{Cov}(x1, x2) = \text{Cov}(x2, x1) = (1 * -1 + 0 * 1 + -1 * 0)/3 = -1/3$

$$\Sigma = \begin{bmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{bmatrix}$$

Find the eigenvalues of Σ:

$\det(\Sigma - \lambda I) = 0$
$(2/3 - \lambda)^2 - (-1/3)^2 = 0$
$(2/3 - \lambda)^2 = 1/9$
$2/3 - \lambda = \pm 1/3$
$\lambda = 2/3 \pm 1/3 = 1, 1/3$

To find the first principal component, take the higher eigenvalue (1):

$(\Sigma - I)x = 0$

$(2/3 - 1)x_1 - 1/3x_2 = 0$
$-1/3x_1 + (2/3 - 1)/x_2 = 0$

$-1/3x_1 - 1/3x_2 = 0$
$-1/3x_1 - 1/3x_2 = 0$

x1 = -x2

First principal component = $[\sqrt{2}, -\sqrt{2}]$
($\sqrt{2}$ chosen so that the magnitude is 1)

3. Nonparametric Tests (35 points)

(a) Parks service claims that the median annual attendance for national parks in the United States is at least 39,000. A random sample of 125 parks reveals that the annual attendances for 79 parks were less than 39,000, the annual attendances for 42 parks were more than 39,000, and the annual attendances for 4 parks were 39,000. At $\alpha = 0.01$, is there enough evidence to reject the park service's claim? (15 points)

H_0: median >= 39,000
H_1: median < 39,000

Test value: $z = 2 * (x + 0.5 - n/2) / \sqrt{n} = 2 * (42 + 0.5 - 121/2) / \sqrt{121} = -3.27$
Critical value (from z table): z = -2.33

z < C.V., so reject the null hypothesis. Yes, there is enough evidence to reject the claim that median attendance is at least 39,000.

(b) Following table shows incomes (in thousands of dollars) of a random sample of 10 female and 10 male software engineers. At $\alpha = 0.10$, can you conclude that there is a difference between female and male incomes? (15 points)

FM	100	86	77	93	101	85	99	100	84	98
M	86	102	89	92	112	88	99	114	99	105

Income	77	84	85	86	86	88	89	92	93	98
Gender	FM	FM	FM	FM	M	M	M	M	FM	FM
Rank	1	2	3	4.5	4.5	6	7	8	9	10

Income	99	99	99	100	100	101	102	105	112	114
Gender	FM	M	M	FM	FM	FM	M	M	M	M
Rank	12	12	12	14.5	14.5	16	17	18	19	20

H_0: $median_M - median_{FM} = 0$
H_1: $median_M - median_{FM} \neq 0$

$n_1 = n_2 = 10$
$\mu_R = n_1(n_1 + n_2 + 1)/2 = 10 * 21/2 = 105$
$\sigma_R = sqrt(n_1n_2(n_1 + n_2 + 1) / 12) = sqrt(100 * 21 / 12) = 13.23$
$R = 4.5 + 6 + 7 + 8 + 12 + 12 + 17 + 18 + 19 + 20 = 123.5$

$z = (R - \mu_R)/\sigma_R = (123.5 - 105)/13.23 = 1.40$
C.V. = ±1.65

-1.65 < 1.40 < 1.65, so the test value does not fall into the critical region and we are not able to reject the null hypothesis that there is no difference.

(c) Write advantages and disadvantages of nonparametric tests (5 points)

Advantages: more general (don't require a certain distribution), generally easier to compute, can be used with categorical data
Disadvantages: waste information (convert quantitative data to qualitative), usually require more data than parametric tests

4. Bonus (4 %)
(a) In simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$, suppose that $E(\epsilon) \neq 0$. By letting $\alpha_0 = E(\epsilon)$, show the this model can always be rewritten with the same slope but a new intercept and error, where the new error has expected zero value. (10 points)

Let $\beta_0' = \beta_0 + \alpha_0$, and ϵ' be the residuals from the resulting model $y = \beta_0' + \beta_1 x$.

$E(\epsilon') = \Sigma_i(y_i - (\beta_0' + \beta_1 x_i))/n = \Sigma_i(y_i - (\beta_0 + \alpha_0 + \beta_1 x_i))/n = \Sigma_i(y_i - (\beta_0 + \beta_1 x_i))/n - \Sigma_i\alpha_0/n$
$= E(\epsilon) - \alpha_0 = \alpha_0 - \alpha_0 = 0$

(b) Show that the error variance is not constant when response variable is binary (10 points)
See HW3 solutions, Q3b.

(c) A medical instrument is used to measure tumor size (x) to determine if the cancer spreads (y = yes/no). Based on some sample data the coefficients for logistic regression are computed as: $\beta_0 = -2$ and $\beta_1 = 0.5$. Compute y for x = 1, 3, 5, 7, 9 and determine labels (yes/no) for respective measurements. (10 points).

$y(x) = 1/(1 + \exp(-\beta_0 - \beta_1 x))$

x	y(x)	Label
1	0.182	No
3	0.378	No
5	0.622	Yes
7	0.818	Yes
9	0.924	Yes

(d) (From Regression question [1C]) Construct a 95% prediction interval for x = 3.5. Write down your conclusion. (10 points)

$t_{0.025} = 2.31$
$s_{est} = 138.26$ (from 1c5)

Prediction interval = $y' \pm t_{0.025} s_{est} * \text{sqrt}(1 + 1/n + n(x - \overline{x})^2/(n\Sigma x^2 - (\Sigma x)^2))$

$= (102.29 + 196.15 * 3.5) \pm$
$2.31 * 138.26 * \text{sqrt}(1 + 1/10 + 10(3.5 - 23.1/10)^2/(10 * 67.35 - 23.1^2))$
$= 788.82 \pm 350.04$

$438.16 < y < 1138.24$

The true y value for x = 3.5 lies in the interval [438.16, 1138.24] with probability of 95%.