MATH 3424 Regression Analysis

Chapter 2 : The Simple Linear Regression Model

2.1 Consider the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, 2, \ldots, n$. Consider the least squares *residuals*, given by $y_i - \hat{y}_i$ ($i = 1, 2, \ldots, n$). Show that

$$\sum_{i=1}^{n} \frac{\hat{y}_i}{n} = \bar{y}$$

2.2 Show that, for the simple linear regression model,

(a) $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$ (b) $\sum_{i=1}^{n} (y_i - \hat{y}_i) x_i = 0$

2.3 Show that the estimator of the error variance σ^2 given by

$$S^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(n-2)}$$

is unbiased; i.e., prove $E(S^2) = \sigma^2$.

2.5 The regression sum of squares for the simple linear regression model, $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ for $i = 1, 2, \ldots, n$, is given by

$$SS_{\text{Reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Show that $E(SS_{\text{Reg}}) = E(MS_{\text{Reg}}) = \sigma^2 + \beta_1^2 S_{xx}$.

2.6 Consider the set of data (y_i, x_i), $i = 1, 2, \ldots, n$, and the following two candidate models

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ($i = 1, 2, \ldots, n$) (Model A)

$y_i = \gamma_0 + \gamma_1 x_i + \gamma_2 x_i^2 + \varepsilon_i$ ($i = 1, 2, \ldots, n$) (Model B)

Suppose both are fit to the same data. Show $SS_{\text{Res},A} \geq SS_{\text{Res},B}$. (*Hint*: Consider model A as a rival to model B. Which of the following will result in the smallest residual *SS*? Model B with $\gamma_0 = b_0$, $\gamma_1 = b_1$, $\gamma_2 = 0$ (b_0 and b_1 are estimated from model A); or Model B with γ_0, γ_1 and γ_2 replaced by the least squares estimators.)

2.7 There is a controversy in the accounting literature concerning the relationship between accounting rates on stockes and market returns. Theory might suggest a positive, perhaps linear relationship between the two variables. Fifty-four companies were chosen with the regressor varianle (x) as the mean yearly accounting rate for the period 1959 to 1974. The dependent variable (y) is the corresponding mean market rate. The data are given in Table 2.1. For the accounting return data of Table 2.1, compute a *t*-statistic for testing

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

where β_1 is the slope of the regression model.

Table 2.1 Accounting rates and market rates from 1959 to 1974

Company	Market Rate	Accounting Rate
McDonnell Dougles	17.73	17.96
NCR	4.54	8.11
Honeywell	3.96	12.46
TRW	8.12	14.70
Raytheon	6.78	11.90
W.R. Grace	9.69	9.67
Ford Motors	12.37	13.35
Textron	15.88	16.11
Lockheed Aircraft	-1.34	6.78
Getty Oil	18.09	9.41
Atlantic Richfield	17.17	8.96
Radio Corporation of America	6.78	14.17
Westinghouse Electric	4.74	9.12
Johnson & Johnson	23.02	14.23
Champion International	7.68	10.43
R.J. Reynolds	14.32	19.74
General Dynamics	-1.63	6.42
Colgate – Palmolive	16.51	12.16
Coca-Cola	17.53	23.19
International Business Machines	12.69	19.20
Allied Chemical	4.66	10.76
Uniroyal	3.67	8.49
Greyhound	10.49	17.70
Cities Service	10.00	9.10
Philip Morris	21.90	17.47
General Motors	5.86	18.45
Philips Petroleum	10.81	10.06
FMC	5.71	13.30
Caterpillar Tractor	13.38	17.66
Georgia Pacific	13.43	14.59
Minnesota Mining & Manufacturing	10.00	20.94
Standard Oil (Ohio)	16.66	9.62
American Brands	9.40	16.32
Aluminum Company of America	0.24	8.19
General Electric	4.37	15.74
General Tire	3.11	12.02
Borden	6.63	11.44
American Home Products	14.73	32.58
Standard Oil (California)	6.15	11.89
International Paper	5.96	10.06
National Steel	6.30	9.60
Republic Steel	0.68	7.41
Warner Lambert	12.22	19.88
U.S. Steel	0.90	6.97
Bethlehem Steel	2.35	7.90
Armco Steel	5.03	9.34
Texaco	6.13	15.40
Shell Oil	6.58	11.95
Standard Oil (Indiana)	14.26	9.56
Owens Illinois	2.60	10.05
Gulf Oil	4.97	12.11
Tenneco	6.65	11.53
Inland Steel	4.25	9.92
Kraft	7.30	12.27

2.8 Consider the zero intercept model given by

$$y_i = \beta_1 x_i + \varepsilon_i \qquad (i = 1, 2, \ldots, n)$$

with the ε_i normal, independent, with variance σ^2. Let x_0 denote an arbitrary value of x. Show that the 100(1 – α)% confidence interval on $E(y \mid x_0)$ is given by

$$b_1 x_0 \pm t_{\alpha/2,n-1} S \sqrt{\frac{x_0^2}{\sum_{i=1}^n x_i^2}}$$

where $S = \sqrt{\sum_{i=1}^n (y_i - b_1 x_i)^2 / (n-1)}$ and $b_1 = \frac{\sum_{i=1}^n y_i x_i}{\sum_{i=1}^n x_i^2}$.

2.9 An experiment was conducted to study the mass of a tracer material exchanged between the main flow of an open channel and the "dead zone" caused by a sudden open channel expansion. Researchers need this information to improve the water quality modeling capability of a river. It is important to determine the exchange constant K for varying flow conditions. The value of K describes the exchange process when a dead zone appears. In this study, values of the Froude Numbers (N_F) were used to predict K. Froude Numbers are functions of upstream channel velocity and water depth. The data are as follows:

Observation	N_F	K
1	0.012500	-0.12562
2	0.023750	-0.12062
3	0.025625	-0.09625
4	0.030000	-0.08062
5	0.033125	-0.07937
6	0.038125	-0.07312
7	0.038125	-0.07250
8	0.038125	-0.07187
9	0.041250	-0.09375
10	0.043125	-0.05312
11	0.045000	-0.05750
12	0.046875	-0.05750
13	0.047500	-0.02000
14	0.050000	-0.04375
15	0.051250	-0.02000
16	0.056250	-0.05125
17	0.062500	-0.03125
18	0.068750	-0.04875
19	0.077500	-0.01687
20	0.044000	-0.04625

The negative sign of the K values indicate "flushing", the direction of mass transfer out of the dead zone.

(a) Use the least squares procedure to fit the model

$$K_i = \beta_0 + \beta_1 (N_F)_i + \varepsilon_i$$

(b) Compute R^2, S^2 and confidence limits on the mean response at the data locations.

(c) Compute the residuals at each data point.

2.10 Physical fitness testing is an important aspect of athletic training. A common measure of the magnitude of cardiovascular fitness is the maximum volume of oxygen uptake during a strenuous exercise. A study was conducted on 24 middle-aged men to study the influence of the time that it takes to complete a 2-mile run. The oxygen uptake measure was accomplished with standard laboratory methods as the subjects performed on a motor driven treadmill. The data are as follows:

Subject	Maximum Volume of O_2 (y)	Time in Seconds (x)
1	42.33	918
2	53.10	805
3	42.08	892
4	50.06	962
5	42.45	968

Subject	Maximum Volume of O_2 (y)	Time in Seconds (x)
6	42.46	907
7	47.82	770
8	49.92	743
9	36.23	1045
10	49.66	810
11	41.49	927
12	46.17	813
13	48.18	858
14	43.21	860
15	51.81	760
16	53.28	747
17	53.29	743
18	47.18	803
19	56.91	683
20	47.80	844
21	48.65	755
22	53.69	700
23	60.62	748
24	56.73	775

(a) Estimate the parameters of a simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (i = 1, 2, \ldots, 24)$$

(b) Does the time it takes to run a distance of two miles have a significant influence on maximum oxygen uptake? Use

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

to answer the question.

(c) Find 95% confidence intervals on the mean max volume of O_2 for the following values of x, the time to run 2 miles.

(i) $x = 750$; (ii) $x = 775$; (iii) $x = 800$; (iv) $x = 825$; (v) $x = 850$.

2.11 In an experiment to determine the influence of certain physical measures on the performance of punters in American football, 13 punters were used as subjects in an experiment in which the average distance on 10 punts was measured. In addition, measures of left leg and right leg strength (lb lifted) were taken via a weight lifting test. The following data were taken. All subjects use their right legs for punting.

Subject	Left Leg (lb)	Right Leg (lb)	Average Punting Distance
1	170	170	162 ft 6 in.
2	130	140	144 ft 0 in.
3	170	180	147 ft 6 in.
4	160	160	163 ft 6 in.
5	150	170	192 ft 0 in.
6	150	150	171 ft 9 in.
7	180	170	162 ft 0 in.
8	110	110	104 ft 10 in.
9	110	120	105 ft 8 in.
10	120	130	117 ft 7 in.
11	140	120	140 ft 3 in.
12	130	140	150 ft 2 in.
13	150	160	165 ft 2in.

(a) Fit a simple linear regression with punting distance as the response and right leg strength as the independent or regressor variable.

(b) Fit a simple linear regression with left leg strength as the regressor variable.

2.12 Consider the Naval manpower data of Table 2.5.

(a) Fit a regression model containing an intercept.

(b) Compute R^2 (intercept model), $R^2_{(0)}$ for the zero intercept model.

(c) Compute S^2 for both the intercept and the zero intercept model.

(d) Compute the confidence limits on $E(y \mid x_i)$ at the 22 sample installations. Perform the computations for both the intercept and zero intercept models.

(e) Use the information in (a) – (d) to choose between the intercept model and zero intercept model.

Table 2.5 Monthly man-hours expended as a function of items processed

Installation	Items Processed, *x*	Monthly Man-hours, *y*
1	15	85
2	25	125
3	57	203
4	67	293
5	197	763
6	166	639
7	162	673
8	131	499
9	158	657
10	241	939
11	399	1546
12	527	2158
13	533	2182
14	563	2302
15	563	2202
16	932	3678
17	986	3894
18	1021	4034
19	1643	6622
20	1985	7890
21	1640	6610
22	2143	8522

2.13 Discuss the computation of a $100(1 - \alpha)\%$ confidence interval on the slope for the case of the zero intercept model.

2.15 The study "Development of LIFTEST, A Dynamic Technique to Assess Individual Capability of Lift Material" was conducted at Virginia Polytechnic Institute and State University in 1982 to determine if certain static arm strength measures have influence on the "dynamic lift" characteristics of an individual. Twenty-five individuals were subjected to strength tests and then were asked to perform a weight-lifting test in which weight was dynamically lifted overhead. The data are as follows:

Individual	Arm Strength, *x*	Dynamic Lift, *y*
1	17.3	71.4
2	19.3	48.3
3	19.5	88.3
4	19.7	75.0
5	22.9	91.7
6	23.1	100.0
7	26.4	73.3
8	26.8	65.0
9	27.6	75.0
10	28.1	88.3
11	28.2	68.3
12	28.7	96.7
13	29.0	76.7
14	29.6	78.3
15	29.9	60.0

Individual	Arm Strength, x	Dynamic Lift, y
16	29.9	71.7
17	30.3	85.0
18	31.3	85.0
19	36.0	88.3
20	39.5	100.0
21	40.4	100.0
22	44.3	100.0
23	44.6	91.7
24	50.4	100.0
25	55.9	71.1

(a) Estimate, with least squares, the linear regression line.

(b) It is hypothesized that $\beta_0 = 0$, $\beta_1 = 2.2$. Test the appropriate joint hypothesis.

(c) Plot the studentized residuals against x.

2.16 Observations on the yield of a chemical reaction taken at various temperatures were recorded as follows:

x (^{o}C)	y (%)
150	77.4
150	76.7
150	78.2
200	84.1
200	84.5
200	83.7
250	88.9
250	89.2
250	89.7
300	94.8
300	94.7
300	95.9

(a) Fit a simple linear regression, estimating β_0 and β_1.

(b) Compute 95% confidence intervals on $E(y|x)$ at the 4 levels of temperature in the data.

2.18 In the manufacture of commercial wood products, it becomes important to estimate the relationship between the density of a wood product and its stiffiness. A relatively new type of particleboard is being considered, which can be formed with considerably more ease than the accepted commercial product. It is necessary to known at what density the stiffiness of the product compares to the well-known, well-documented commercial product. Thirty particleboards were produced at densities ranging from roughly 8 to 26 pounds per cubic foot, and the stiffiness was measured in pounds per square inch. Table 2.3 shows the data. Construct a normal probability plot, and comment regarding the normal error assumption.

Table 2.3 Density and stiffness for 30 particleboards

Density (x) lb / ft^3	Stiffness (y) lb / in.^{2}
9.50	14814.00
8.40	17502.00
9.80	14007.00
11.00	19443.00
8.30	7573.00
9.90	14191.00
8.60	9714.00
6.40	8076.00
7.00	5304.00
8.20	10728.00
17.40	43243.00
15.00	25319.00
15.20	28028.00
16.40	41792.00
16.70	49499.00

Density (x) lb / ft^3	Stiffness (y) lb / in.^{2}
15.40	25312.00
15.00	26222.00
14.50	22148.00
14.80	26751.00
13.60	18036.00
25.60	96305.00
23.40	104170.00
24.40	72594.00
23.30	49512.00
19.50	32207.00
21.20	48218.00
22.80	70453.00
21.70	47661.00
19.80	38138.00
21.30	53045.00

2.19 An entomological experiment was conducted at Virginia Polytechnic Institute and State University to study the survivability of stalk borer larvae. Nine chambers were used to simulate different conditions. It was of interest to develop a model relating the mean size of larvae as a function of the stalk head diameter.

Chamber	Head Diameter (cm), x	Size of Larvae (cm), y
1	333	2.274
	624	2.765
	1102	2.912
	1352	2.831
	1643	2.859
2	333	2.081
	624	2.726
	1102	2.869
	1352	2.881
	1643	2.814
3	397	2.129
	744	2.725
	1314	2.823
	1609	2.875
	1950	2.921
4	397	2.533
	744	2.823
	1314	2.921
	1609	2.873
	1950	2.903
5	461	2.281
	864	2.669
	1526	2.730
	1871	2.818
	2274	2.767
6	461	2.133
	864	2.577
	1526	2.692
	1871	2.832
	2274	2.769
7	295	2.056
	558	2.617
	831	2.780
	1070	2.793
	1362	2.871
8	237	1.407
	444	2.148
	784	2.570
	962	2.661
	1169	2.757
9	109	1.352
	204	1.603
	360	1.929
	442	2.103
	537	2.180

(a) Estimate the linear regression line that relates larvae size to stalk head diameter.

(b) Plot the residuals against x, the head diameter. Comment on appropriateness of the simple linear regression model. Can you comment on the effect of the different chamber conditions?

2.20 Consider the fixed zero intercept model. The appropriate estimator of σ^2 is given by

$$S^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(n-1)}$$

Show that S^2 is unbiased for σ^2.

2.21 Suppose the experimenter postulates a model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad (i = 1, 2, \ldots, n)$$

where β_0 is known.

(a) What is the appropriate least squares estimator of β_1 ? Justify your answer.

(b) What is the variance of the estimator of the slope in (a) ?

(c) Write out an expression for the confidence interval on $E(y|x)$ for this model.

2.22 Consider the studentized residuals

$$\frac{y_i - \hat{y}_i}{S\sqrt{1-(1/n)-[(x_i-\bar{x})^2/S_{xx}]}}$$

The denominator is found by merely constructing the variance of $y_i - \hat{y}_i$, namely

$$Var(y_i - \hat{y}_i) = \sigma^2\left[1 - \frac{1}{n} - \frac{(x_i-\bar{x})^2}{S_{xx}}\right]$$

and then standardizing $y_i - \hat{y}_i$.

(a) Show that

$$\sum_{i=1}^{n} \frac{Var(y_i - \hat{y}_i)}{\sigma^2} = n - 2$$

(b) Under the conditions that the ε_i are i.i.d. $N(0, \sigma^2)$, does the studentized residual have a t-distribution with $n - 2$ degrees of freedom? If not, why not?

2.23 Consider a situation in which the regression data set is divided into two parts as follows.

x	y	
x_1	y_1	Portion 1
x_2	y_2	
$\vdots$	$\vdots$	
x_{n1}	y_{n1}	
x_{n_1+1}	y_{n_1+1}	Portion 2
$\vdots$	$\vdots$	
$x_{n_1+n_2}$	$y_{n_1+n_2}$	

The model is given by

$$y_i = \beta_0^{(1)} + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, \ldots, n_1$$

$$= \beta_0^{(2)} + \beta_1 x_i + \varepsilon_i \qquad i = n_1 + 1, \ldots, n_1 + n_2$$

In other words there are two regression lines with common slope. Using the centered model,

$$y_i = \beta_0^{(1)*} + \beta_1 (x_i - \bar{x}_1) + \varepsilon_i \qquad i = 1, 2, \ldots, n_1$$

$$= \beta_0^{(2)*} + \beta_1 (x_i - \bar{x}_2) + \varepsilon_i \qquad i = n_1 + 1, \ldots, n_1 + n_2$$

where $\bar{x}_1 = \sum_{i=1}^{n_1} \frac{x_i}{n_1}$ $\bar{x}_2 = \sum_{i=n_1+1}^{n_1+n_2} \frac{x_i}{n_2}$

Show that the least squares estimate of β_1 is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_1) y_i + \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x}_2) y_i}{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2 + \sum_{i=n_1+1}^{n_1+n_2} (x_i - \bar{x}_2)^2}$$

(*Hint*: Set up $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ according to this centered model. Then find estimators for $\beta_0^{(1)*}$ $\beta_0^{(2)*}$, and β_1 that minimize $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.)

2.25 Consider the simple linear regression model, $y_i = \beta_0^* + \beta_1 (x_i - \bar{x}) + \varepsilon_i$ for $i = 1, 2, \ldots, n$, show that the least squares slope is given by

$$b_1 + \beta_1 + \sum_{i=1}^{n} d_i \varepsilon_i \qquad \text{where} \qquad d_i = \frac{(x_i - \bar{x})}{S_{xx}}$$

Also show that

$$b_0 = \beta_0 + \bar{\varepsilon} - \bar{x} \sum_{i=1}^{n} d_i \varepsilon_i$$