THE UNIVERSITY OF BRITISH COLUMBIA
FINAL EXAMINATION, April 2016
STAT 306, Finding Relationships in Data

TIME: 2 1/2 hours

THIS EXAMINATION CONSISTS OF 4 PAGES. THERE ARE 6 PROBLEMS FOR A TOTAL OF 100 POINTS. PLEASE CHECK TO ENSURE THAT THIS PAPER IS COMPLETE.

Special Instructions: Please write your name and student number on the front page of the exam booklet. A non-programmable calculator and a two-sided sheet of notes are allowed. The problems can be done in any order. READ THE QUESTIONS CAREFULLY.
Suggestion: use the left-hand side of booklet for scratch work, and the right-hand side for submitted answers.
For a large-sample z or t critical value for a 95% confidence interval, use 1.96 or 2. For smaller degrees of freedom, $\mathrm{t}_{35,0.975} = 2.030$, $\mathrm{t}_{24,0.975} = 2.064$, $\mathrm{t}_{21,0.975} = 2.080$, $\mathrm{t}_{10,0.975} = 2.228$, $\mathrm{t}_{3,0.975} = 3.182$.

(16) 1. Scientists are assessing if the CO2 concentration at a single location can be used to estimate the mean global annual temperature. The variables are x=CO2 concentration (in parts per 10^8) at top of Mauna Loa in Hawaii and y=mean annual temperature in Celsius (over land and water across the globe). Consider the data for 1959–1995 ($n = 37$ years); for example $(x, y) = (2.452, 16.53)$ in 1959 and $(x, y) = (6.412, 16.83)$ in 1995. Summary statistics are: $\overline{x} = 4.639$, $s_x = 1.226$, $\overline{y} = 16.621$, $s_y = 0.098$, $r_{xy} = 0.508$.

(a) What are the slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ of the least squares regression line for y as a function of x?

(b) What is the residual SD : $\hat{\sigma} = \sqrt{(n-2)^{-1}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$?

(c) Suppose (x_i, y_i) are indexed by increasing year. Describe a residual plot to check if the residuals are serially correlated.

(d) Assuming that the residual plot (c) is acceptable, what is a prediction interval for mean temperature of the globe when the CO2 concentration is 6.5 parts per 10^8?

(16) 2. An experiment was done to find the rate of bottle return for 6 different deposit levels from 2 to 30 cents. Data and logistic regression output are shown below.

deposit	2	5	10	20	25	30
#sold	100	100	100	100	100	100
#returned	14	20	34	59	81	90

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.10829    0.19107  -11.03   <2e-16
deposit      0.13715    0.01072   12.79   <2e-16
    Null deviance: 831.75  on 599  degrees of freedom
Residual deviance: 609.41  on 598  degrees of freedom,
AIC: 613.41
```

(a) Obtain the sample proportions for the 6 deposit levels. Compare with the fitted proportions using the estimated β's given above. Is the logistic model a good fit?

(b) Assuming the logistic model is reasonable, what is an approximate 95% confidence interval for the slope parameter β_1?

(c) What is the estimated probability that the bottle will be returned if the deposit is 15 cents?

(d) What is the estimated deposit level so that the expected return rate is 50%?

(18) 3. http://www.cdc.gov/nchs/nhanes.htm has data sets from the National Health and Nutrition Examination Survey. Some variables such as total (body) fat are based on dual-energy x-ray absorptiometry, so are harder to measure. Other body measurements such as weight (wt), height (ht), body mass index (bmi), waist circumference, triceps skin fold (tri) are easier to measure. So one might try to find a prediction equation for total fat (tofat) based on these other body measurements. Below are some summary statistics for a sample of 631 men between the ages of 20 and 40 in the 2003–2004 survey. The right-hand side has the sample correlation matrix.

summary	*wt*	*ht*	*bmi*	*waist*	*tri*	*tofat*
Min.	42.80	154.1	16.01	63.70	3.50	6.785
1st-Qu.	70.60	170.8	23.36	84.30	9.30	15.854
Median	79.90	175.8	26.11	92.30	13.10	20.777
3rd-Qu.	91.50	180.9	29.41	100.45	18.40	26.079
Max.	132.40	193.0	41.97	138.30	38.80	55.052
Mean	81.75	175.5	26.51	92.82	14.19	21.622
SD	15.23	7.46	4.47	12.22	6.54	8.097

cor	*wt*	*ht*	*bmi*	*waist*	*tri*	*tofat*
wt	1.000	0.407	0.891	0.885	0.610	0.880
ht	0.407	1.000	−0.046	0.109	0.010	0.149
bmi	0.891	−0.046	1.000	0.916	0.665	0.887
waist	0.885	0.109	0.916	1.000	0.705	0.942
tri	0.610	0.010	0.665	0.705	1.000	0.795
tofat	0.880	0.149	0.887	0.942	0.795	1.000

Next are some summaries of the best fitting 2-, 3-, 4- and 5-variable models from `regsubsets()` in R.

```
 2-variable  Estimate    SE    tvalue |  3-variable  Estimate    SE    tvalue
(Intercept) -29.657    0.803  -36.94 | (Intercept) -26.985    0.787  -34.28
waist         0.503    0.010   48.25 | wt            0.121    0.012   10.29
tri           0.321    0.019   16.49 | waist         0.366    0.016   22.30
                                     | tri           0.329    0.018   18.23

4-variable   Estimate  SE      tvalue | 5-variable  Estimate SE      tvalue
(Intercept) -24.373    2.886  -8.45   | (Intercept) 20.794   12.322   1.69
wt            0.131    0.016   8.26   | wt           0.408    0.075   5.43
ht           -0.014    0.015  -0.94   | ht          -0.271    0.070  -3.88
waist         0.357    0.019  18.47   | bmi         -0.892    0.237  -3.77
tri           0.327    0.018  18.06   | waist        0.367    0.019  19.00
                                      | tri          0.329    0.018  18.33
```

Other summaries of the above fits are:

#vars	residSD	df	R^2	adjR^2	C_p	CVRMSE
2	2.268	628	0.9218	0.9216	123.3	2.277
3	2.099	627	0.9331	0.9328	17.1	2.111
4	2.099	626	0.9332	0.9328	18.2	2.113
5	2.077	625	0.9347	0.9342	6.0	2.094

(a) Which of the 5 explanatory variables leads to the best 1-variable simple linear regression model? Why?

(b) What is the adjusted R^2 value for the 1-variable model in (a)?

(c) What is the partial correlation $r_{\text{tofat,tri;waist}}$?

(d) Consider the partial correlation $r_{\text{tofat,ht;wt,bmi,waist,tri}}$. Which of the following is correct based on the above summaries: (i) $r_{\text{tofat,ht;wt,bmi,waist,tri}} > 0$; (ii) $r_{\text{tofat,ht;wt,bmi,waist,tri}} < 0$; (iii) insufficient information to determine the sign. Explain your choice.

(e) Which of the four models would you prefer? Explain your reasoning in one sentence.

(f) Why does the coefficient for `wt` change so much among the models with 3,4,5 explanatory variables?

(21) 4. A paper helicopter experiment (http://www.paperhelicopterexperiment.com/) was run to find some optimal dimensions. Explanatory variables are body length and body width (both in cm) of a piece of paper before the folding/cutting is done to produce the helicopter. The response variable y is the flight time (to land on floor) after release from a height of 2.5 m. With variables abbreviated as `len` and `wid`, the data are:

len	wid	flighttime	len	wid	flighttime	len	wid	flighttime	len	wid	flighttime
5.6	1.4	1.98	6.6	1.4	1.85	6.1	2.1	1.82	6.6	2.1	2.00
6.1	1.4	1.89	6.1	1.4	1.87	6.1	0.7	1.43	5.6	0.7	1.25
6.6	2.1	2.08	6.6	2.1	1.86	6.1	1.4	1.73	6.6	0.7	1.00
6.1	0.7	1.41	5.6	0.7	1.10	6.6	0.7	0.93	6.1	2.1	1.85
5.6	1.4	1.87	5.6	2.1	1.69	5.6	1.4	1.83	5.6	0.7	1.21
6.1	0.7	1.30	5.6	2.1	1.69	6.6	1.4	1.98	5.6	2.1	1.61
6.6	0.7	0.93	6.6	1.4	1.94	6.1	2.1	1.94			

Summaries from fitting a linear equation, a quadratic with original variables and a quadratic with centred variables (clen=len-6, cwid=wid-1.5) are given below.

```
fit1        Estimate Std.Error tvalue Pr(>|t|)
(Intercept)  0.736    0.702    1.049   0.305
len          0.038    0.113    0.334   0.742
wid          0.475    0.081    5.866   5e-06 ***
Residual SD: 0.2403 on 24 df; Multiple R2: 0.5899, Adjusted R2:  0.5557

fit2       Estimate SE   tvalue Pr(>|t|)  | fit2c    Estimate  SE   tvalue    pval
Intercept -10.994 6.722 -1.635 0.117      | Intercept   1.973 0.046 42.720 2e-16 ***
len         4.042 2.200  1.837 0.080 .    | clen        0.152 0.064  2.376 0.027 *
wid         0.231 0.612  0.376 0.710      | cwid        0.282 0.042  6.641 1e-06 ***
I(len^2)   -0.373 0.180 -2.074 0.051 .    | I(clen^2)  -0.373 0.180 -2.074 0.051 .
I(wid^2)   -0.769 0.092 -8.370 4e-08 *** | I(cwid^2)  -0.769 0.092 -8.370 4e-08 ***
len:wid     0.393 0.091  4.321 .0003 *** | clen:cwid   0.393 0.091  4.321 .0003 ***
Residual SD : 0.1102 on 21 df            | Residual SD: 0.1102 on 21 df
Multiple R2: 0.9245, AdjustedR2: 0.9065  | Multiple R2:  0.9245, AdjustedR2: 0.9065
```

$(\mathbf{X}^T\mathbf{X})^{-1}$ for the third regression (labelled `fit2c`) is given below.

$$\begin{array}{rrrrrr}
0.1755 & 0.0607 & -0.0319 & -0.4178 & -0.2198 & -0.0068 \\
0.0607 & 0.3357 & -0.0068 & -0.5333 & -0.0000 & 0.0680 \\
-0.0319 & -0.0068 & 0.1479 & 0.0000 & 0.1388 & -0.0680 \\
-0.4178 & -0.5333 & 0.0000 & 2.6667 & 0.0000 & 0.0000 \\
-0.2198 & -0.0000 & 0.1388 & 0.0000 & 0.6942 & 0.0000 \\
-0.0068 & 0.0680 & -0.0680 & 0.0000 & 0.0000 & 0.6803
\end{array}$$

(a) What are three things in the regression summaries that indicate that the quadratic function fit is better than the linear function?

(b) For the quadratic model, which β's are invariant to the shifting of length and width by constants?

(c) What are the best residual plots to check that the quadratic model adequately handles the curved surface for flight time as a function of length and width?

(d) For the $\mathbf{X}$ matrix for quadratic with centred variables, what are the values of the 6 columns in row 1?

(e) Using the above summaries, obtain a point estimate, prediction SE and a prediction interval from the quadratic fit when length=6 cm, width=1.5 cm or clen=0, cwid=0. (Note: little arithmetic is needed).

(f) Based on the fitted quadratic, what are the estimated optimal length and width to maximize flight time?

(14) 5. Consider a data set to daily log returns of 8 stocks in the American market in 2010–2011 (504 trading days); the stocks are: LO (Lorillard Tobacco), MO (Altria Tobacco), PM (Philip Morris Tobacco), RAI (Reynolds Tobacco), DPS (Dr Pepper Snapple Beverage), KO (Coca-Cola), MNST (Monster Beverage), PEP (Pepsico). Principal component analysis was applied to both the sample covariance matrix and sample correlation matrix. A summary of some results are below.

stock	LO	MO	PM	RAI	DPS	KO	MNST	PEP
sample mean	0.0009	0.0010	0.0011	0.0011	0.0007	0.0005	0.0017	0.0003
sample SD	0.0149	0.0100	0.0125	0.0119	0.0158	0.0104	0.0202	0.0101
row 1	-0.0272	0.0006	-0.0088	0.0014	-0.0116	-0.0122	0.0041	0.0120

```
            sample covariance matrix  |   sample correlation matrix
Importance of components:
          Comp1  Comp2  Comp3  Comp4  |  Comp1  Comp2  Comp3  Comp4
SD        0.028  0.016  0.013  0.010  |  2.102  0.913  0.832  0.802
PropofVar 0.515  0.175  0.117  0.071  |  0.552  0.104  0.087  0.080
CumProp   0.515  0.690  0.807  0.878  |  0.552  0.657  0.743  0.824

Loadings:  sample covariance matrix  |   sample correlation matrix
          Comp1  Comp2  Comp3  Comp4 |  Comp1  Comp2  Comp3  Comp4
LO       -0.363 -0.389  0.383  0.689 | -0.331 -0.504        -0.401
MO       -0.274 -0.173        -0.139 | -0.395 -0.216
PM       -0.337 -0.206  0.187 -0.347 | -0.381 -0.194 -0.102  0.163
RAI      -0.335 -0.218  0.145 -0.122 | -0.400 -0.250
DPS      -0.362 -0.174 -0.879  0.223 | -0.285  0.469  0.600 -0.544
KO       -0.278 -0.102        -0.402 | -0.384  0.133         0.455
MNST     -0.546  0.823         0.126 | -0.271  0.536 -0.749 -0.257
PEP      -0.246 -0.105        -0.378 | -0.356  0.271  0.255  0.483
```

(a) How many components are needed to explain 80% of the variation in the data using the sample covariance matrix?

(b) Interpret the first two components from the coefficients of the loadings for both sets of output.

(c) The first observation in the data set is given in the above table along with the sample means and SDs of the 8 variables. For observation 1, what is the value of comp1 for the left-hand side (in the new coordinate system centred at the vector of sample means). Write down an expression without doing the calculation.

(15) 6. Let data be (x_i, y_i), $i = 1, \ldots, n$, where all of the x_i are positive. For simple linear regression with model $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, the least squares estimate of β_1 is $\hat{\beta}_1 = \sum_{i=1}^{n} a_i y_i$ where $a_i = (x_i - \overline{x})/[(n-1)s_x^2]$, and the least squares intercept is $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$. Let $\hat{B}_1$ be the least squares slope when considered as a random variable. Determine the variance of $\hat{B}_1$ under the two scenarios given below.

(a) (heteroscedastic) ϵ_i are independent $N(0, \sigma_i^2)$ random variables, where $\sigma_i^2 = \gamma_0 + \gamma_1 x_i$ with $\gamma_0 > 0$ and $\gamma_1 > 0$ being parameters. What is $\mathrm{Var}\,(Y_i)$? What is $\mathrm{Var}\,(\hat{B}_1)$?

(b) (serial dependence with i an index for a time sequence) ϵ_i are serially dependent $N(0, \sigma^2)$ random variables such that $\mathrm{Cov}\,(\epsilon_i, \epsilon_j) = \sigma^2 \gamma^{|i-j|}$ for $i, j \in \{1, \ldots, n\}$, where $0 < \gamma < 1$. What is $\mathrm{Var}\,(\hat{B}_1)$?

(c) Assume the model in (b), what is the correlation of ϵ_1, ϵ_2 (equivalently, the correlation of $\epsilon_i, \epsilon_{i+1}$)?

(d) [Harder.] Assume the model in (b). Let $e_i = y_i - \hat{y}_i$ and $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $i = 1, \ldots, n$. Based on your answers in (b) and (c), suggest an estimated standard error of $\hat{\beta}_1$ that is a function of the e_i and x_i.