Midterm II—Statistics 232A

5 December 2013

NAME: Solutions SID: ____________

Instructions: Closed Book. Solve the problems in the blank space below each. Justify your answers cogently. You may consult two, double-sided, letter-size sheets of personal notes and a hand calculator.

PROBLEM 1. Consider the **general linear model**

$$y = Cm + e, \tag{1}$$

where y and e are $n \times 1$, m is $p \times 1$, C is $n \times p$ with rank $p < n$, and the components of e are i.i.d. with mean 0, nonzero finite variance σ^2, and finite fourth moment. Both $m = (m_1, m_2, \ldots, m_p)'$ in R^p and σ^2 are unknown.

Suppose $0 \leq d \leq p - 1$ is fixed. Let $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_d)'$ and let the $\{x_i : 1 \leq i \leq p\}$ be the distinct values of a single covariate. Define

$$m_i(d) = \sum_{k=0}^{d} \gamma_k x_i^k \qquad 1 \leq i \leq p, \tag{2}$$

and let $m(d) = (m_1(d), m_2(d), \ldots m_p(d))'$. The **$d$-th degree polynomial submodel** specifies that $y = Cm(d) + e$ for some unknown $\gamma \in R^{d+1}$. The distribution of e is unchanged.

a) (4 points) Using notation in (1) and (2), write the d-th degree polynomial submodel as a standard linear model in matrix form. Identify the regression parameter vector and the design matrix explicitly. State the rank of the latter.

Let $\underset{p\times(d+1)}{W(d)} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^d \\ 1 & x_2 & \cdots & x_2^d \\ \vdots & & & \vdots \\ 1 & x_p & \cdots & x_p^d \end{pmatrix}$ $\quad \underset{(d+1)\times 1}{\gamma} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_d \end{pmatrix}$ Then $m(d) = W(d)\gamma$

Submodel: $y = X(d)\gamma + e$ where $\underset{n\times(d+1)}{X(d)} = C\,W(d)$.

Design matrix is $X(d)$

Regression parameters are γ

$\text{rank}(X(d)) = \text{rank}(W(d)) = d+1$

Indeed $\text{rank}(W(d)) = \text{rank}[(C'C)^{-1}C'CW(d)] \leq \text{rank}(CW(d)) \leq \text{rank}(W(d))$

and $\text{rank}(W(d)) = d+1$ by the fundamental theorem of algebra (cf. class material)

b) (4 points) Do any of the polynomial submodels (2) coincide with the general linear model (1)? Justify your answer.

$W(p-1)$ is $p \times p$ and $\text{rank}(W(p-1)) = (p-1)+1 = p$

$\mathcal{R}(X(p-1)) = \mathcal{R}(CW(p-1)) \subset \mathcal{R}(C) = \mathcal{R}[CW(p-1)W^{-1}(p-1)] \subset \mathcal{R}(CW(p-1))$
$= \mathcal{R}(X(p-1))$

Hence $\mathcal{R}(X(p-1)) = \mathcal{R}(C)$

so submodel $y = X(p-1)\gamma + e$ coincides with the general model

Define the **risk** of any estimator $\tilde{m}$ of m to be

$$R(\tilde{m}, m, \sigma^2) = p^{-1}\mathrm{E}|C\tilde{m} - Cm|^2, \tag{3}$$

the expectation being computed under general model (1). Let $\hat{m}$ denote the least squares estimator of m and σ^2 respectively in model (1). Let $\hat{m}(d)$ denote the least squares estimator of $m(d)$ in submodel (2).

c) (2 points) Give algebraically an unbiased estimator $\hat{R}(\hat{m}(d))$ for the risk $R(\hat{m}(d), m, \sigma^2)$. Give in simplest algebraic form an unbiased estimator $\hat{R}(\hat{m})$ for the risk $R(\hat{m}, m, \sigma^2$. Express both estimated risks as functions of $|y - C\hat{m}|^2$, $|y - C\hat{m}(d)|^2$, n, p, and d.

Let $r(d) = \text{rank}\,(X(d)) = d+1$ and $\hat{\sigma}^2 = (n-p)^{-1}|y - C\hat{m}|^2$

Then, by class results

$\hat{R}(\hat{m}(d)) = p^{-1}[|y - C\hat{m}(d)|^2 + (2r(d) - n)\hat{\sigma}^2]$

$\hat{R}(\hat{m}) = \hat{R}(\hat{m}(p-1)) = p^{-1}[|y - C\hat{m}|^2 + \hat{\sigma}^2(2p-n)]$ using $\hat{m} = \hat{m}(p-1)$

$= p^{-1}[\hat{\sigma}^2(n-p) + \hat{\sigma}^2(2p-n)] = \hat{\sigma}^2$

d) (2 points) Give the customary F-statistic for testing, at level α, the null hypothesis that y satisfies the polynomial submodel of degree d versus the alternative that it does not but the general model still holds. Express your answer in terms of $|y - C\hat{m}|^2$, $|y - C\hat{m}(d)|^2$, n, p and d. State the distribution of this test statistic under the null hypothesis.

The F-statistic is $T(d) = \dfrac{[|y - C\hat{m}(d)|^2 - |y - C\hat{m}|^2]/(p - r(d))}{\hat{\sigma}^2}$

with $\hat{\sigma}^2 = (n-p)^{-1}|y - C\hat{m}|^2$

The null distribution is F with $p - r(d)$ and $n-p$ d.f.

The Canadian earnings data records observations on the logarithm of income versus age for a sample of 205 persons. The ages of these persons range over every age from 21 to 65 years. See the handout for plots of the data and some least squares fits.

In applying the foregoing theory, take y to be the 205 observed log(incomes). The distinct covariate values are $x_i = i + 20$, where $1 \le i \le 45$. Here m_i and $m_i(d)$ are, respectively, the mean log(income) at age x_i under the general model (1) and under the d-th degree polynomial submodel (2). It is found numerically that

$$|y - C\hat{m}|^2 = 47.26 \tag{4}$$

and that

$$|y - C\hat{m}(2)|^2 = 63.54, \quad |y - C\hat{m}(3)|^2 = 61.98, \quad |y - C\hat{m}(4)|^2 = 56.58, \quad |y - C\hat{m}(5)|^2 = 55.69. \tag{5}$$

SID: ____________________

e) (10 points) For the Canadian earnings data, compute the estimated risk $\hat{R}(\hat{m}(d))$ for $d = 2, 3, 4, 5$ and the estimated risk $\hat{R}(\hat{m})$. Report your findings. What may you conclude?

$p = 45, \quad n = 205, \quad n - p = 160, \quad r(d) = d + 1, \quad \hat{\sigma}^2 = \frac{47.26}{160} = .2954$

$r(2) = 3, \quad \hat{R}(\hat{m}(2)) = \frac{63.54 - (6 - 205)(.2954)}{45} = \frac{63.54 - 58.78}{45} = \frac{4.76}{45} = \boxed{.1057}$

$r(3) = 4, \quad \hat{R}(\hat{m}(3)) = \frac{61.98 - (8 - 205)(.2954)}{45} = \frac{61.98 - 58.19}{45} = \frac{3.79}{45} = \boxed{.0842}$

$r(4) = 5, \quad \hat{R}(\hat{m}(4)) = \frac{56.58 - (10 - 205)(.2954)}{45} = \frac{56.58 - 57.60}{45} = \frac{-1.02}{45} = \boxed{-.0227}$

$r(5) = 6, \quad \hat{R}(\hat{m}(5)) = \frac{55.69 - (12 - 205)(.2954)}{45} = \frac{55.69 - 57.01}{45} = \frac{-1.32}{45} = \boxed{-.0293}$

$\hat{R}(\hat{m}) = \hat{\sigma}^2 = \boxed{.2954}$

The candidate estimator with smallest estimated risk is $\hat{m}(5)$

f) (8 points) For the Canadian earnings data and for $d = 2, 3, 4, 5$, test at level $\alpha = .05$ the null hypothesis that that y satisfies the polynomial submodel of degrees d versus the alternative that it does not but the general model still holds. Report for each d the numerical value of the F-test statistic, the pertinent degrees of freedom, and the outcome of the test. See the handout for a table of critical values. What may you conclude?

$T(2) = \frac{(63.54 - 47.26)/42}{.2954} = \frac{.3876}{.2954} = \boxed{1.31}$ $\quad p - r(2) = 45 - 3 = 42$, d.f. are 42 and 160

$T(3) = \frac{(61.98 - 47.26)/41}{.2954} = \frac{.3590}{.2954} = \boxed{1.22}$ $\quad p - r(3) = 45 - 4 = 41$, df are 41 and 160

$T(4) = \frac{(56.58 - 47.26)/40}{.2954} = \frac{.2330}{.2954} = \boxed{.789}$ $\quad p - r(4) = 45 - 5 = 40$, df are 40 and 160

$T(5) = \frac{(55.69 - 47.26)/39}{.2954} = \frac{.2162}{.2954} = \boxed{.732}$ $\quad p - r(5) = 45 - 6 = 39$, df are 39 and 160

From the F-table, the critical value for $\alpha = .05 \approx 1.50$

None of the tests rejects the null hypothesis at level .05

i.e. there is insufficient evidence against any of these low degree polynomial fits.

Testing differs from estimation!

SID: ____________________

PROBLEM 2. Consider the standard **general linear model** $y = X\beta + e$. Here y is $n \times 1$ and X is $n \times p$ with $p < n$. The components of the error vector e are independent, identically distributed with mean 0 and finite unknown variance σ^2.

Consider the **submodel** $y = X_0\beta_0 + e$, where $\mathcal{R}(X_0) \subset \mathcal{R}(X)$ and e is as above. Let $r_0 = \operatorname{rank}(X_0)$. Least squares theory for this submodel yields the variance estimator

$$\hat{\sigma}_0^2 = |y - X_0X_0^+y|^2/(n - r_0).$$

a) (5 points) Show that, under the general model,

$$\mathrm{E}(\hat{\sigma}_0^2) = \sigma^2 + |X\beta - X_0X_0^+X\beta|^2/(n - r_0).$$

Let $\eta = X\beta$ so $y = \eta + e$. Let $A = X_0X_0^+$. Then A is symmetric, idempotent with $\operatorname{rank}(A) = \operatorname{tr}(A) = \operatorname{rank}(X_0) = r_0$ (Lab # 1)

$$E|y - Ay|^2 = E|(I_n - A)\eta + (I_n - A)e|^2 = |\eta - A\eta|^2 + E\operatorname{tr}[(I_n - A)ee'(I_n - A)] + 2E[\eta'(I_n - A)^2 e]$$

$$= |\eta - A\eta|^2 + \operatorname{tr}[(I_n - A)^2 \sigma^2 I_n] = |\eta - A\eta|^2 + \operatorname{tr}(I_n - A)\sigma^2$$

$$= |\eta - A\eta|^2 + (n - \operatorname{tr}(A))\sigma^2 = |\eta - A\eta|^2 + (n - r_0)\sigma^2$$

Hence $E(\hat{\sigma}_0^2) = \dfrac{E|y - Ay|^2}{n - r_0} = \sigma^2 + \dfrac{|\eta - A\eta|^2}{n - r_0} = \sigma^2 + \dfrac{|X\beta - X_0X_0^+X\beta|^2}{n - r_0}$

b) (5 points) Suppose further that $X'X = I_p$ and $X_0 = XP$, where P is a $p \times p$ symmetric, idempotent matrix. Show that

$$\mathrm{E}(\hat{\sigma}_0^2) = \sigma^2 + |\beta - P\beta|^2/(n - \operatorname{tr}(P)).$$

$r_0 = \operatorname{rank}(X_0) = \operatorname{rank}(XP) = \operatorname{rank}(P) = \operatorname{tr}(P)$

using $\operatorname{rank}(P) = \operatorname{rank}(X'XP) \le \operatorname{rank}(XP) \le \operatorname{rank}(P)$ and Lab # 1

$$X_0X_0^+ = X_0(X_0'X_0)^+X_0' = XP[P\underbrace{X'X}_{I_p}P]^+PX' = XP(P^2)^+PX' = XPX'$$

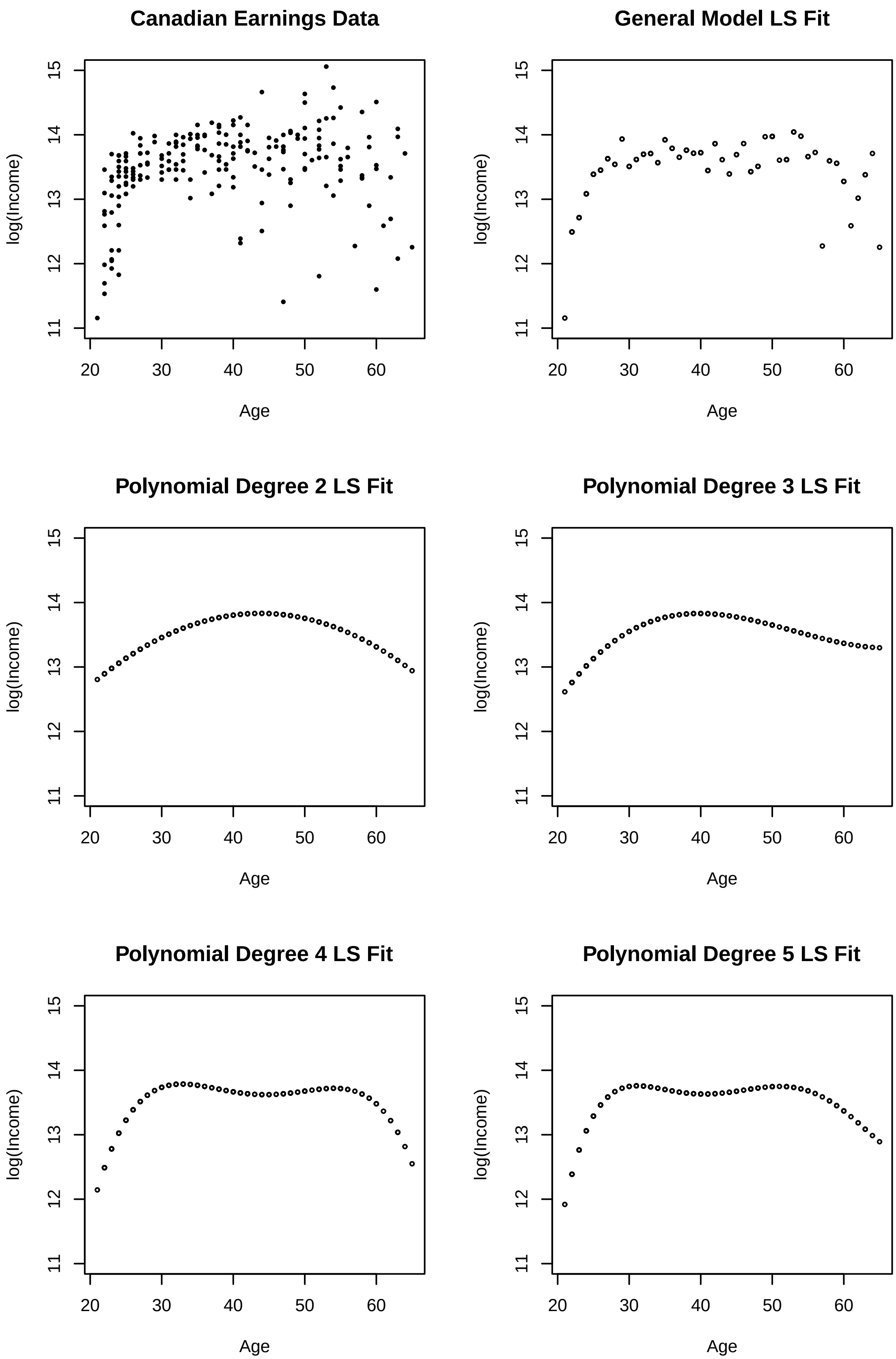
because $P^2 = P$, $P^+ = P$

Hence $X_0X_0^+X\beta = XPX' \cdot X\beta = XP\beta$ and so

$$|X\beta - X_0X_0^+X\beta|^2 = |X\beta - XP\beta|^2 = |X(\beta - P\beta)|^2$$

$$= (\beta - P\beta)'\underbrace{X'X}_{I_p}(\beta - P\beta) = |\beta - P\beta|^2$$

Using part a, $E(\hat{\sigma}_0^2) = \sigma^2 + \dfrac{|\beta - P\beta|^2}{n - \operatorname{tr}(P)}$



Canadian Earnings Data
General Model LS Fit
Polynomial Degree 2 LS Fit
Polynomial Degree 3 LS Fit
Polynomial Degree 4 LS Fit
Polynomial Degree 5 LS Fit
log(Income)
Age
11
12
13
14
15
20
30
40
50
60

F Values for $\alpha = 0.05$

d_2 \ d_1	10	12	15	20	24	30	40	60	120	inf
1	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	19.4	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.5
3	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	1.91	1.83	1.75	1.66	1.10	1.55	1.50	1.43	1.35	1.25
inf	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00