Theory Exam 2

36-401, Fall 2015, Section B

17 November 2015

Record all of your answers in the blue-book provided; if you need more space, ask for another blue-book. Show work for all problems; even a completely correct answer will receive no credit if unsupported by work.

The only permitted electronic device for this exam is a calculator. If you do not have a calculator, you may use a calculator app on a phone, but you may not use any other functionality of the phone, and it must be in airplane mode.

You are allowed notes covering both sides of one 8.5×11 inch piece of paper.

1. You are given a dataset with one response variable Y, and two predictor variables X_1, X_2, with $n = 5$ observations. You are to fit the following multiple linear regression model, *without an intercept*:

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$$

   (a) (5) Write out the matrix form of the multiple linear regression, including the error assumptions as well. Indicate the dimensions for all matrices.

   (b) (5) Express $\frac{1}{n}\mathbf{X}^T\mathbf{X}$ in terms of sums related to X_{i1}, X_{i2}, Y_i etc.

   (c) (10) To estimate the coefficients, we minimize the sum of squares $\sum_{i=1}^{n}(Y_i - \beta_1 X_{i1} - \beta_2 X_{i2})^2$ and derive the following estimating equations:

$$\begin{aligned} \sum X_{i1} Y_i &= \beta_1 \sum X_{i1}^2 + \beta_2 \sum X_{i1} X_{i2} \\ \sum X_{i2} Y_i &= \beta_1 \sum X_{i1} X_{i2} + \beta_2 \sum X_{i2}^2 \end{aligned}$$

   Re-write the minimization problem using matrices, derive the matrix form of the estimating equations, and show that it has the solution

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

   (d) (10) Show that $\widehat{\beta}$ is an unbiased estimate of β and derive the formula of the variance-covariance matrix of $\widehat{\beta}$

   (e) (5) What is the dimension of the hat matrix $\mathbf{H}$ for this model? Show that $\mathbf{H}$ is idempotent and symmetric.

   (f) (5) Express the formula of residual vector $\mathbf{e}$ in terms of the hat matrix. Derive the expectation and variance-covariance matrix of $\mathbf{e}$ in matrix notation.

2. Recall the SENIC dataset that you worked with in previous homework assignments. The primary objective of the original study was to determine factors that are associated with the average estimated probability of acquiring infection in hospitals. Someone suggested to use a multivariate linear regression model to study the relationship between the infection risk and the following predictor variables.

Variable Name	Description
Risk	Average estimated probability of acquiring infection in hospital (in percent)
Age	Average age of patients (in years)
Length	Average length of stay of all patients in hospital (in days)
Region	Geographic region (NE, NC, S, or W)
Nurses	Average number of full-time-equivalent registered and licensed nurses
DailyPatient	Average number of patients in hospital per day during study period
MedSchool	Binary: is the hospital affiliated with a medical school?

Assume that all the usual model assumptions are satisfied, including Gaussian noise. Refer to the R output on the next two pages, and the tables at the end of the exam (as needed).

(a) (2) Find n, the number of observations, and p, the number of parameters.

(b) (2) Find the maximum likelihood estimate of σ^2.

(c) (2) Interpret the estimated intercept of the model. Is there anything strange about this value? (Explain.)

(d) *Nurses*

i. (4) Interpret the estimated coefficient on the number of nurses.

ii. (5) Find and interpret a 90% confidence interval for β_{Nurses}.

iii. (5) Test, with $\alpha = 0.05$, whether or not the true $\beta_{Nurses} = 0$. State your conclusion in the context of the problem.

(e) (5) Interpret the estimated coefficients for geographic regions.

(f) (5) Someone suggests that the region shouldn't matter. Can you test whether region is a statistically significant predictor, using this output? If yes, perform a suitable test and state your conclusions. If not, explain why not, and what additional information you would need.

(g) (5) Interpret the coefficient for the average number of patients in hospital per day. *Hint:* think carefully about interactions.

(h) (5) Test whether or not the coefficients for being affiliated to a medical school, for the average number of patients and for their interaction are all simultaneously zero, at $\alpha = 0.05$.

(i) (5) Someone suggests including the interaction of region and whether or not the hospital is affiliated with a medical school. How many variables will be added into our model? How will you interpret their coefficients?

```
##
## Call:
## lm(formula = Risk ~ Age + Length + Region + Nurses + MedSchool *
##     DailyPatient, data = SENIC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.31830 -0.75079  0.04736  0.63598  2.71323
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.0930793  1.3872010   0.788   0.4325
## Age                    -0.0245642  0.0241755  -1.016   0.3120
## Length                  0.4141846  0.0739120   5.604 1.76e-07
## RegionNC                0.1240157  0.2987682   0.415   0.6789
## RegionS                -0.1213014  0.3019914  -0.402   0.6888
## RegionW                 0.8229193  0.3887624   2.117   0.0367
## Nurses                  0.0037485  0.0018902   1.983   0.0500
## MedSchool               0.9523789  0.7746822   1.229   0.2217
## DailyPatient           -0.0005447  0.0020460  -0.266   0.7906
## MedSchool:DailyPatient -0.0036388  0.0019785  -1.839   0.0688
##
## Residual standard error: 1.07 on 103 degrees of freedom
## Multiple R-squared:  0.4146,Adjusted R-squared:  0.3634
## F-statistic: 8.104 on 9 and 103 DF,  p-value: 5.23e-09
## Analysis of Variance Table
##
## Model 1: Risk ~ Age + Length + Nurses + MedSchool * DailyPatient
## Model 2: Risk ~ Age + Length + Region + Nurses + MedSchool * DailyPatient
##   Res.Df    RSS Df Sum of Sq
## 1    106 126.80
## 2    103 117.89  3    8.9098
```

```
## Analysis of Variance Table
##
## Model 1: Risk ~ Age + Length + Region + Nurses
## Model 2: Risk ~ Age + Length + Region + Nurses + MedSchool * DailyPatient
##   Res.Df    RSS Df Sum of Sq
## 1    106 124.81
## 2    103 117.89  3     6.913
```

```
## Analysis of Variance Table
##
## Model 1: Risk ~ Age + Length + Nurses
## Model 2: Risk ~ Age + Length + Region + Nurses + MedSchool * DailyPatient
##   Res.Df    RSS Df Sum of Sq
## 1    109 132.91
## 2    103 117.89  6    15.019
```

3. We say that one model is "strictly larger" than another, or that the first model "nests" the second, when the first model includes all the terms in the second, and more besides.

   (a) (3) Explain why the MSE of the strictly larger model is never more than that of the smaller model (when estimated on the same data).

   (b) (7) Suppose we compare models according to the C_p criterion. Express the condition under which we will prefer a strictly larger model, in terms of an inequality relating the difference in MSEs to some combination of the number of parameters, the sample size, and $\hat{\sigma}^2$.

   (c) (5) Suppose we compare two models on the same data set, but neither one is strictly larger than the other. Give an example of two such models, referring to the SENIC data set from the previous problem for definiteness. Does your result from (3b) still hold, even though the models are not nested? (Explain.)

	1	2	3	4	5	6	7
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10
101	3.94	3.09	2.69	2.46	2.30	2.19	2.10
102	3.93	3.09	2.69	2.46	2.30	2.19	2.10
103	3.93	3.08	2.69	2.46	2.30	2.19	2.10
104	3.93	3.08	2.69	2.46	2.30	2.19	2.10
105	3.93	3.08	2.69	2.46	2.30	2.19	2.10
106	3.93	3.08	2.69	2.46	2.30	2.19	2.10
107	3.93	3.08	2.69	2.46	2.30	2.18	2.10
108	3.93	3.08	2.69	2.46	2.30	2.18	2.10
109	3.93	3.08	2.69	2.45	2.30	2.18	2.09
110	3.93	3.08	2.69	2.45	2.30	2.18	2.09
111	3.93	3.08	2.69	2.45	2.30	2.18	2.09
112	3.93	3.08	2.69	2.45	2.30	2.18	2.09
113	3.93	3.08	2.68	2.45	2.29	2.18	2.09
ChiSq	3.84	5.99	7.81	9.49	11.10	12.60	14.10

Table 1: 95th percentile of the F distribution, for selected numbers of degrees of freedom. Columns give the first number of degrees of freedom (numerator of F statistic), rows the second (denominator). The last row gives the 95th percentile of the χ^2 distributions.

	0.005	0.025	0.05	0.1	0.9	0.95	0.975	0.995
100	-2.63	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.63
101	-2.63	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.63
102	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
103	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
104	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
105	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
106	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
107	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
108	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
109	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
110	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
111	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
112	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
113	-2.62	-1.98	-1.66	-1.29	1.29	1.66	1.98	2.62
Inf	-2.58	-1.96	-1.64	-1.28	1.28	1.64	1.96	2.58

Table 2: Selected quantiles of t distributions, with selected degrees of freedom; the last row gives quantiles of the z distribution.