Practice final exam, STATS 401 W18

Name:

UMID:

Instructions. You have a time allowance of 120 minutes. The exam is closed book and closed notes. Any electronic devices in your possession must be turned off and remain in a bag on the floor. If you need extra paper, please number the pages and put your name and UMID on each page.

You may use the following formulas. Proper use of these formulas may involve making appropriate definitions of the necessary quantities.

(1) $\mathbf{b} = \big(\mathbb{X}^\top \mathbb{X}\big)^{-1} \mathbb{X}^\top \mathbf{y}$

(2) $\mathrm{Var}(\hat{\beta}) = \sigma^2 \big(\mathbb{X}^\top \mathbb{X}\big)^{-1}$

(3) $\mathrm{Var}(\mathbb{A}\mathbf{Y}) = \mathbb{A}\mathrm{Var}(\mathbf{Y})\mathbb{A}^\top$

(4) $\mathrm{Var}(X) = \mathrm{E}\big[(X - \mathrm{E}[X])^2\big] = \mathrm{E}[X^2] - \big(\mathrm{E}[X]\big)^2$

(5) $\mathrm{Cov}(X, Y) = \mathrm{E}\big[\big(X - \mathrm{E}[X]\big)\big(Y - \mathrm{E}[Y]\big)\big] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$

(6) The binomial (n, p) distribution has mean np and variance $np(1-p)$.

(7) $f = \dfrac{(\mathrm{RSS}_0 - \mathrm{RSS}_a)/(q-p)}{\mathrm{RSS}_a/(n-q)}.$

From `?pnorm`:

```
pnorm(q, mean = 0, sd = 1)
qnorm(p, mean = 0, sd = 1)
q: vector of quantiles.
p: vector of probabilities.
```

Questions 1–4 refer to data on 113 hospitals from the Study on the Efficacy of Nosocomial Infection Control (SENIC), provided in the R dataframe `senic`. The primary purpose of this study is to look for properties of hospitals associated with high (or low) rates of hospital-acquired infections, which have the technical name of *nosocomial infections*. The rate of nosocomial infections is measured by the variable `Infection risk`. The variables are described as follows:

Hospital: index from 1 to 113

Length of stay: average duration (in days) for all patients

Age: average age (in years) for all patients

Infection risk: estimated percentage of patients acquiring an infection in hospital

Culture: average number of cultures for each patient without signs or symptoms of hospital-acquired infection, times 100

X-ray: number of X-ray procedures divided by number of patients without signs or symptoms of pneumonia, times 100

Beds: average number of beds in the hospital

Med school: does the hospital have an affiliated medical school (1=Yes;2=No)

Region: geographic region (1=North-East, 2=North-Central, 3=South, 4=West)

Patients: average daily census of number of patients in the hospital

Nurses: average number of full-time equivalent registed and licensed nurses

Facilities: percent of 35 specific facilities and services which are provided by the hospital

Throughout Questions 1–4, we will write y_i for the measured infection risk in hospital i for $i = 1, \ldots, n$ with $n = 113$. We will consider sample models of the form $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$ where $\mathbf{y} = (y_1, \ldots, y_n)$, and $\mathbf{e} = (e_1, \ldots, e_n)$ is a vector of residual error. The design matrix $\mathbb{X}$ will be different in various models. You may use this notation without explanation, but other additional notation you use should be defined as appropriate.

```
head(senic)
```

```
##   Infection.risk Length.of.stay Culture X.ray Region
## 1            4.1           7.13     9.0  39.6      4
## 2            1.6           8.82     3.8  51.7      2
## 3            2.7           8.34     8.1  74.0      3
## 4            5.6           8.95    18.9 122.8      4
## 5            5.7          11.20    34.5  88.9      1
## 6            5.1           9.76    21.9  97.0      2
```

1. Factors and their coding in R. Consider the following two models, specified in R code as

```
lm1 <- lm(Infection.risk~Region, data=senic)
lm2 <- lm(Infection.risk~factor(Region), data=senic)
```

Write down the first six rows of the design matrix for each of `lm1` and `lm2`. Which model makes more sense to use?

2.

3.

4.

5.

6.

7.

8.

Acknowledgments: The SENIC study was described in a sequence of articles in The American Journal

of Epidemiology, Volume 111, Issue 5, 1980, Pages 465–653. The dataset used here comes from Kutner, Nachtscheim, Neter and Li (2005) *Applied Linear Statistical Models*, 5th Edition, Appendix C1.
