Practice final exam, STATS 401 W18

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.

Responses will be assessed on quality of explanation as well as whether they lead to a correct answer.

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

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

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

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

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

(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.
```

All questions in this exam 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 the exam, 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[,c("Infection.risk","Length.of.stay","Culture","Region","Beds")])
```

```
##   Infection.risk Length.of.stay Culture Region Beds
## 1            4.1           7.13     9.0      4  279
## 2            1.6           8.82     3.8      2   80
## 3            2.7           8.34     8.1      3  107
## 4            5.6           8.95    18.9      4  147
## 5            5.7          11.20    34.5      1  180
## 6            5.1           9.76    21.9      2  150
```

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`. Sufficient information to do this is provided in the R output above. Which model makes more sense to use?

Solution. Discussed in class.

2. An F-test. Do different regions tend to different levels of infection risk? To investigate the difference between regions while controlling for some other important explanatory variables, the following model was considered:

```
lm3 <- lm(Infection.risk~Length.of.stay+Culture+factor(Region),data=senic)
anova(lm3)
```

```
## Analysis of Variance Table
##
## Response: Infection.risk
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## Length.of.stay    1  57.305  57.305  60.573 4.759e-12 ***
## Culture           1  33.397  33.397  35.302 3.572e-08 ***
## factor(Region)    3   9.451   3.150   3.330   0.02234 *
## Residuals       107 101.227   0.946
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Solution. Check class notes, lab notes, homeworks and quiz 2 for various examples of writing out hypotheses and test

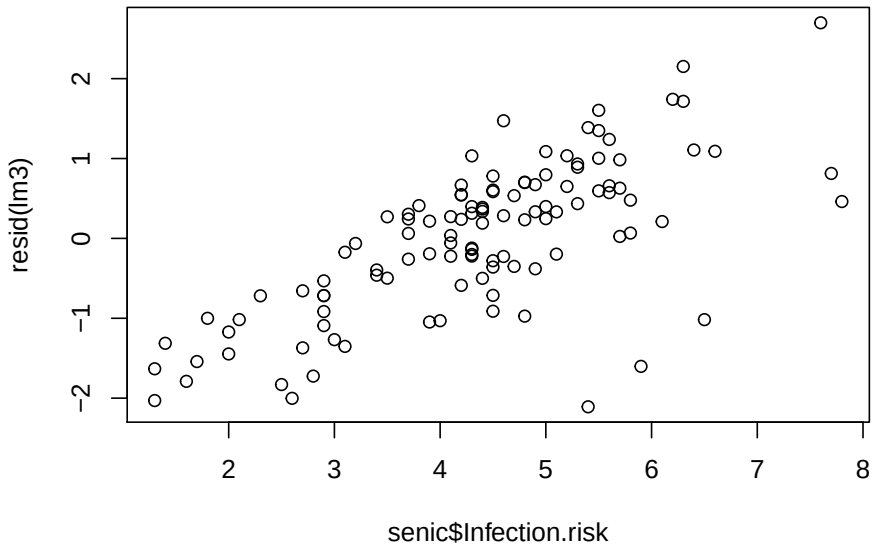
(a) Write out in full the null hypothesis, H_0, for a test of whether there are regional differences in infection risk. You may use either a matrix or subscript form.

(b) Write out the alternative hypothesis, H_a. You do not have to repeat any definitions you have already made in the part (a). You may use either a matrix or subscript form.

(c) Carry out an F test of these hypotheses, giving explanation of how this test is constructed. What do you conclude?

3. Model diagnostics. Here, we investigate what happens when we plot residuals against the response variable.

```
plot(x=senic$Infection.risk,y=resid(lm3))
```



The graph above shows residuals of a model for infection risk plotted against the data. Does this plot indicate a flaw in the model? If so, what might you try to do to fix the model? Explain.

Solution. Discussed in class.

4. Model interpretation. We consider the fitted model represented by the R object `lm3`. For the purposes of this question, we should ignore the possibility that we might have preferred a different model.

```
summary(lm3)
```

```
##
## Call:
## lm(formula = Infection.risk ~ Length.of.stay + Culture + factor(Region),
##     data = senic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1084 -0.6564  0.2108  0.6047  2.6994
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.161945   0.648026  -0.250  0.80314
```

```
## Length.of.stay   0.341202   0.056888   5.998 2.76e-08 ***
## Culture          0.058438   0.009729   6.007 2.65e-08 ***
## factor(Region)2  0.320149   0.265812   1.204  0.23109
## factor(Region)3  0.199331   0.271845   0.733  0.46501
## factor(Region)4  1.030779   0.350042   2.945  0.00397 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9726 on 107 degrees of freedom
## Multiple R-squared:  0.4973, Adjusted R-squared:  0.4738
## F-statistic: 21.17 on 5 and 107 DF,  p-value: 1.147e-14
```

(a) An interpretation of the sample coefficient for `Length.of.stay` is that each additional day in hospital leads to an additional 0.34 percent risk of infection, on average. Comment on the validity of this conclusion.

Solution. This conclusion is the causal interpretation of a regression coefficient. For an observational study, such as this, the causal interpretation is not necessarily valid. There may be reverse causation (more nosocomial infection causes longer stays rather than the direction of causality implied by how we set up our linear model) or confounding variables. Length of stay may be a proxy for some other variable, such as severity of cases that may be more directly responsible for the observed association.

(b) What do you conclude from the R^2 and F statistics presented in the above summary? Which of these statistics is more useful for the purpose of this study? Explain.

Solution. The F statistic corresponds to an F test with all the fitted model parameters against an null hypothesis that the mean is constant. The strong evidence against this null (p-value$= 1.15 \times 10^{-14}$) is firm evidence that there is structure in the data which can be investigated further. The R^2 statistic is equivalent to this F statistic, as shown in the notes. It does not usually come with an associated p-value. It does have more immediate interpretation, that the model explains about 1/2 of the variation in the data. The goal of this study is to identify detectable explanatory variables predicting rates of nosocomial infection. This is more closely related to carrying out hyothesis testing under the null hypothesis that the explanatory variables are irrelevant.

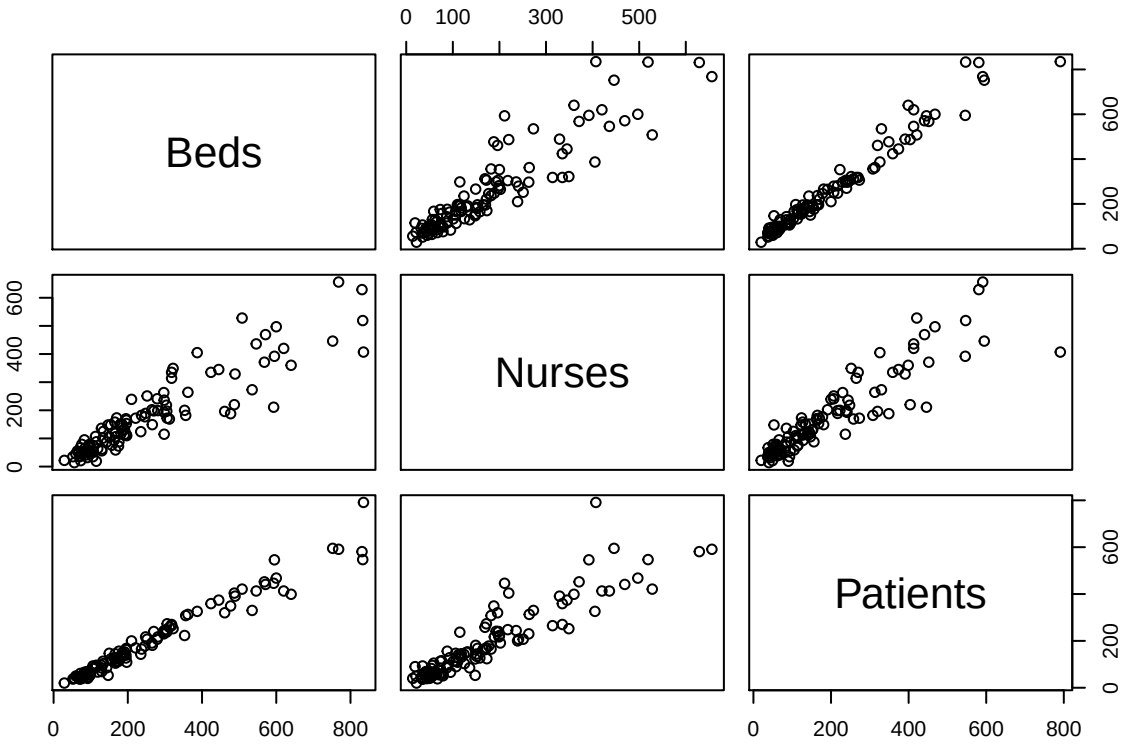
5. Colinearity.

(a) One might suspect that the number of beds, the number of nurses and the number of patients may be colinear. Explain why, in words.

Solution. Discussed in class.

To investigate this, we can make the following scatterplot:

```
pairs(~Beds+Nurses+Patients,data=senic)
```



(b) From this scatterplot, what is the strongest colinearity you can identify? Explain.

Now, let's see the consequences of this colinearity when comparing two fitted models.

```
lm4 <- lm(Infection.risk~Length.of.stay+Culture+factor(Region)+Beds+Nurses,data=senic)
summary(lm4)
```

```
##
## Call:
## lm(formula = Infection.risk ~ Length.of.stay + Culture + factor(Region) +
##     Beds + Nurses, data = senic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.10054 -0.57121  0.08802  0.49542  2.69449
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.0157925  0.6611265  -0.024  0.98099
## Length.of.stay   0.3063009  0.0620664   4.935 3.03e-06 ***
## Culture          0.0548310  0.0096905   5.658 1.34e-07 ***
## factor(Region)2  0.2796140  0.2673065   1.046  0.29794
## factor(Region)3  0.1753505  0.2741104   0.640  0.52376
## factor(Region)4  0.9592797  0.3450850   2.780  0.00645 **
## Beds            -0.0006948  0.0012659  -0.549  0.58428
## Nurses           0.0026102  0.0016876   1.547  0.12495
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.9517 on 105 degrees of freedom
## Multiple R-squared:  0.5277, Adjusted R-squared:  0.4963
## F-statistic: 16.76 on 7 and 105 DF,  p-value: 1.079e-14
```

```
lm5 <- lm(Infection.risk~Length.of.stay+Culture+factor(Region)+Beds,data=senic)
summary(lm5)
```

```
##
## Call:
## lm(formula = Infection.risk ~ Length.of.stay + Culture + factor(Region) +
##     Beds, data = senic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3833 -0.6282  0.1597  0.6121  2.5088
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.1544684  0.6561642   0.235   0.8143
## Length.of.stay  0.2888724  0.0614344   4.702 7.79e-06 ***
## Culture         0.0575578  0.0095911   6.001 2.77e-08 ***
## factor(Region)2 0.2197442  0.2662201   0.825   0.4110
## factor(Region)3 0.1023132  0.2717795   0.376   0.7073
## factor(Region)4 0.9449267  0.3472181   2.721   0.0076 **
## Beds            0.0010895  0.0005247   2.076   0.0403 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9579 on 106 degrees of freedom
## Multiple R-squared:  0.517,  Adjusted R-squared:  0.4896
## F-statistic: 18.91 on 6 and 106 DF,  p-value: 7.308e-15
```

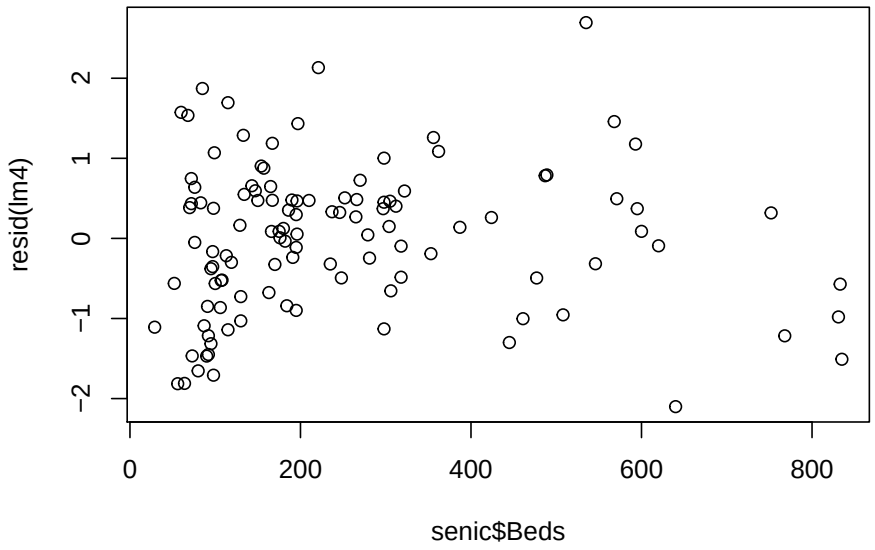
(c) Explain the evidence for colinearity of number of beds and number of nurses from the outcome of fitting these two models.

(d) What might have happened to our investigation of the data if we had started with the model coded as `lm4` and had not considered the possibility of colinearity?

(e) Suppose the sample model for `lm4` is written as $\mathbf{y} = \mathbb{X}\mathbf{b} + \mathbf{e}$ where $\mathbb{X} = [x_{ij}]$ with x_{i1} being the number of beds for hospital i and x_{i2} being the number of nurses. Use output provided to suggest a nonzero vector α such that $\mathbb{X}\boldsymbol{\alpha}$ is close to $\mathbf{0}$. As part of this answer, you will have to figure out the dimensions of $\mathbb{X}$.

6. Nonlinearity. Up to this point, we have considered variables entering the model only as additive effects without interactions. Now, let's do some model diagnostics on the role of hospital size by plotting residuals against number of beds.

```
plot(senic$Beds,resid(lm4))
```



(a) Interpret this residual plot.

Solution. Discussed in class.

Now let's try modeling the nonlinearity by adding a quadratic term, coded in R as follows:

```
lm6 <- lm(Infection.risk~Length.of.stay+Culture+factor(Region)+Beds+I(Beds^2),data=senic)
summary(lm6)
```

```
##
## Call:
## lm(formula = Infection.risk ~ Length.of.stay + Culture + factor(Region) +
##     Beds + I(Beds^2), data = senic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.25780 -0.59384  0.06929  0.53627  2.29886
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3.886e-01  6.365e-01  -0.611 0.542841
## Length.of.stay   2.683e-01  5.827e-02   4.604 1.16e-05 ***
## Culture          5.693e-02  9.058e-03   6.285 7.59e-09 ***
## factor(Region)2  2.146e-01  2.514e-01   0.854 0.395216
## factor(Region)3  9.381e-02  2.566e-01   0.366 0.715437
## factor(Region)4  1.070e+00  3.296e-01   3.248 0.001563 **
## Beds             7.060e-03  1.677e-03   4.210 5.40e-05 ***
## I(Beds^2)       -7.644e-06  2.051e-06  -3.727 0.000314 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9045 on 105 degrees of freedom
## Multiple R-squared:  0.5734, Adjusted R-squared:  0.545
## F-statistic: 20.16 on 7 and 105 DF,  p-value: < 2.2e-16
```

(b) Assess the consequences of adding the quadratic dependence on number of beds. This should involve considering the summary of `lm6` in comparison with `lm4`.

Solution. Discussed in class.

(c) Superficially, it may seem strange that we can add a nonlinear term, such as a quadratic, while continuing to work within the statistical framework of linear models. Explain how to resolve this paradox.

Solution. A linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is linear in the coefficient vector $\boldsymbol{\beta}$. The quadratic term in the explanatory variable leads to a nonlinear function of an explanatory variable in $\mathbb{X}$, but any choice of $\mathbb{X}$ is a valid linear model.

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.
