STAT 151A Final Exam practice questions

Spring 2020

May 6, 2020

Instructions: answer each of the questions below. Please be sure to give clear reasons for your answers when asked and to show work to receive partial credit. You do not need to reduce all numerical calculations to a single number.

Question	Total points
Q1	12
Q2	6
Q3	9
Q4	6
Q5	12
Q6	14
Q7	12
	60

1. Let's imagine that 80 students took a particular course at Berkeley of whom 20 were freshmen, 20 were sophomores, 20 were juniors and 20 were seniors. In R, I have saved the final scores (out of 100) for the 20 freshmen in the vector `g1`, for the 20 sophomores in `g2`, juniors in `g3` and seniors in `g4`. Consider the following output:

```
> mean(g1)
[1] 57.96
> sd(g1)
[1] 3.92
> mean(g2)
[1] 64.13
> sd(g2)
[1] 3.91
> mean(g3)
[1] 67.60
> sd(g3)
[1] 6.92
> mean(g4)
[1] 71.22
> sd(g4)
[1] 5.77
```

Also, for $i = 1, \ldots, 80$, let

- y_i: Final score of the i^{th} student in the class.
- x_{i1}: Takes the value 1 if the i^{th} student is a freshman and 0 otherwise.

- x_{i2}: Takes the value 1 if the i^{th} student is a sophomore and 0 otherwise.
- x_{i3}: Takes the value 1 if the i^{th} student is a junior and 0 otherwise.
- x_{i4}: Takes the value 1 if the i^{th} student is a senior and 0 otherwise.

I fit the linear model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i, i = 1, \ldots, n$$

to this data via R to obtain the following output:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4)

Residuals:
     Min       1Q   Median       3Q      Max
-14.4685  -3.6042  -0.1473   3.1631  12.5674

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    XXXXX       1.18   60.24  < 2e-16 ***
x1            -13.26       1.67   -7.93  1.5e-11 ***
x2             -7.09       1.67   -4.24  6.2e-05 ***
x3             -3.62       1.67   -2.16    0.034 *
x4                NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: XXXXX on 76 degrees of freedom
Multiple R-squared:  0.4734,Adjusted R-squared:  XXXXX
F-statistic: 22.77 on 3 and 76 DF,  p-value: 1.279e-10
```

(a) Why does the R output above say "`1 not defined because of singularities`"? Give reasons for your answer and suggest a way to fix the problem. (**2 points**)

(b) Fill in the 3 missing values in the R output with proper reasoning. (**6 points**).

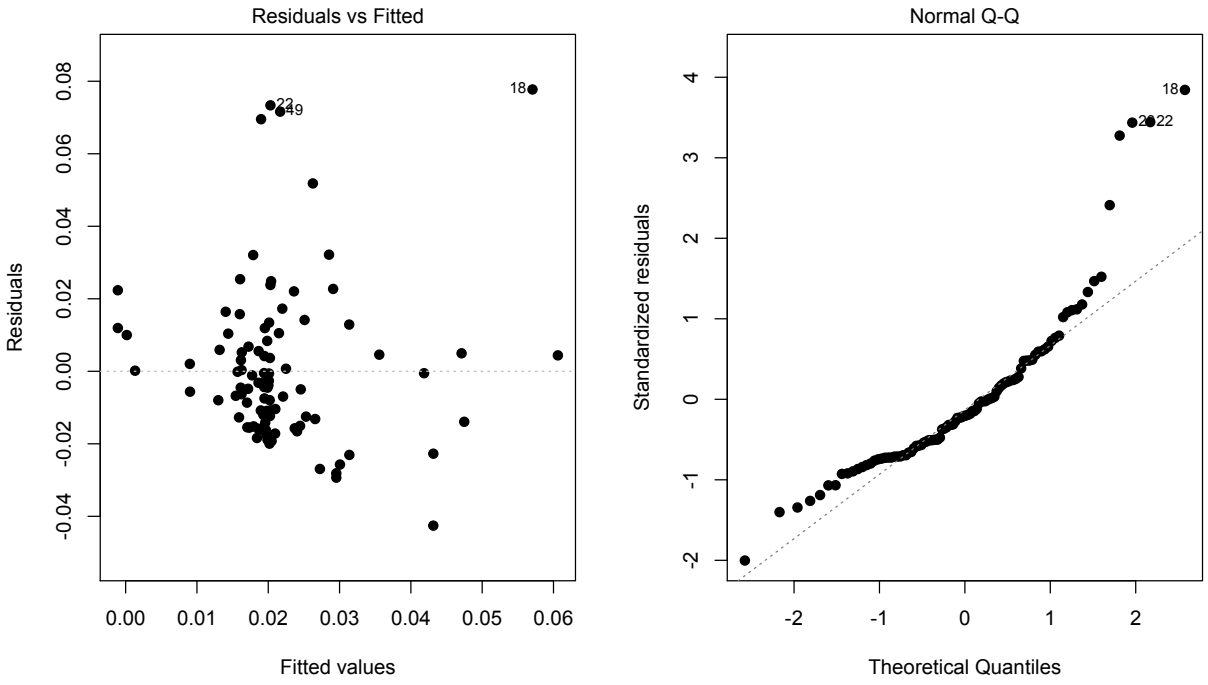
(c) Explain why the standard error estimates for the coefficients of `x1`, `x2`, and `x3` are all the same. (**3 points**).

2. We run a linear regression with the response income and explanatory variables age and ethnicity. Ethnicity takes on four values: "White", "Black", "Hispanic", "Asian"

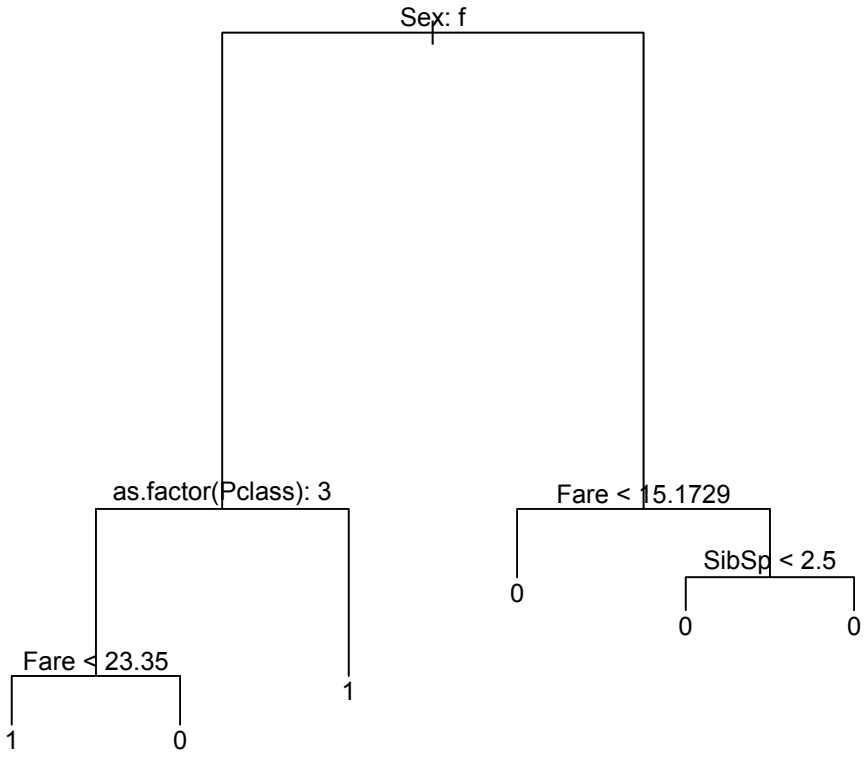
```
>  summary(lm(y~age*ethnicity))

Call:
lm(formula = y ~ age * ethnicity)

Residuals:
     Min       1Q   Median       3Q      Max
-2.12271 -0.69124 -0.02662  0.61052  2.61588

Coefficients:
                       Estimate Std. Error   t value Pr(>|t|)
(Intercept)           5.000e+04  5.884e-01 84971.706  < 2e-16 ***
age                   4.028e-02  1.376e-02     2.928  0.00430 **
ethnicityBlack        8.936e-01  1.201e+00     0.744  0.45886
```

```
ethnicityHispanic      5.705e-01  8.090e-01     0.705  0.48246
ethnicityAsian        -1.612e+00  1.933e+00    -0.834  0.40649
age:ethnicityBlack    -7.941e-02  2.838e-02    -2.798  0.00625 **
age:ethnicityHispanic -7.880e-02  1.895e-02    -4.159 7.16e-05 ***
age:ethnicityAsian     1.176e-02  4.406e-02     0.267  0.79024
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.009 on 92 degrees of freedom
Multiple R-squared: 0.6596,Adjusted R-squared: 0.6337
F-statistic: 25.47 on 7 and 92 DF,  p-value: < 2.2e-16
```

Below are the residual vs. fitted plot and the Q-Q plot from the model. Describe what problems you see, if any, in the assumptions of the model. If you see problems in these diagnostic plots, describe what you might suggest to get an improved regression model. (**6 points**).



3. We revisit a subset of the dataset *titanic* from the homework. The response variable is *Survived* which takes the value 1 if the passenger survived and 0 otherwise. We fit a classification tree to this dataset with the response variable being *Survived* and the explanatory variables being *Pclass* (proxy for the class in which the passenger travelled, with three levels 1, 2 and 3), *Sex* (gender), *SibSp* (number of siblings/spouses aboard), *Parch* (number of Parents/Children aboard), *Fare* (ticket fare) and *Embarked* (port of embarkation; has three levels: C for Cherbourg, Q for Queenstown and S for Southampton). The following plot summarizes the fitted tree.



(a) Based on the fitted tree, do we predict that a female passenger who travelled in *Pclass* 3 with a fare of 15 will survive or not? (**2 points**).

(b) Again based on the tree *rt*, do we predict that a female passenger who travelled in *Pclass* 1 with a fare of 25 will survive or not? (**2 points**).

(c) The leaf of the tree oriented furthest to the right in the plot contains 186 individuals, 64 of whom survived. Calculate the Gini index at this leaf (**3 points**).

(d) The decision rule SibSp < 2.5 at the far right of the plot separates two leaves, but both leaves give the same prediction. Assuming that the tree was grown using the Gini index, explain why this might have occurred (**4 points**).

4. The summary of a fitted model is:

```
Coefficients:
            Estimate Std. Error t value
(Intercept)    1.12      1.00    1.12
x              0.50      0.25    2.0

Residual standard error: 2.2 on 40 degrees of freedom
Multiple R-squared:  0.22,
F-statistic: 2.27 on 1 and 40 DF
```

Below are the percentiles for a bootstrapped estimate of the sampling distribution of the studentized $\hat{\beta}$ in this simple linear model.

0.01	0.015	0.02	0.025	0.03	0.035	.04	0.045	0.05	0.055
-2.3	-2.0	-1.7	-1.6	-1.5	-1.4	-1.3	-1.3	-1.2	-1.2
0.95	0.955	0.95	0.965	0.97	0.975	0.98	0.985	0.99	0.995
1.7	1.7	1.7	1.9	2.0	2.0	2.2	2.4	2.7	3.4

Construct a 95% confidence interval for β, using the bootstrapped studentized distribution. (**6 points**).

5. We have data for 195 new high school students who each must choose one of three curricular tracks: academic, general, or vocational. We fit a multinomial logit model to predict their probabilities for choosing each option, using as explanatory variables a dummy variable for female gender and prior test scores for reading, writing, and math. Here is a summary of the fitted model:

```
Call:
multinom(formula = prog ~ female + read + write + math, data = hschoice)

Coefficients:
         (Intercept) femalefemale     read     write    math
academic       -4.92        0.084  0.02860  0.000294  0.0781
vocation        4.10        0.242 -0.00888 -0.050700 -0.0246

Std. Errors:
         (Intercept) femalefemale   read  write   math
academic        1.43        0.413 0.0269 0.0295 0.0307
vocation        1.55        0.470 0.0305 0.0318 0.0355

Residual Deviance: 337.7572
AIC: XXXXXXXXX
```

   (a) Fill in the missing value in the model output. (**2 points**).

   (b) Rank the predicted probabilities for academic, general, and vocational tracks from largest to smallest for a male student with scores of 40 for reading and writing and a score of 30 for math. (**4 points**).

   (c) Suppose we want to conduct a hypothesis test for whether the dummy variable for female gender is necessary in this model. Write the associated null hypothesis in mathematical notation, describe a test statistic we could use to test it, and give its null distribution. (**4 points**).

   (d) Now suppose we combine the students choosing academic and general tracks and instead model the choice of vocational or non-vocational tracks, using logistic regression with all four explanatory variables in the original model. Compute and the degrees of freedom for this logistic model and compare it to the degrees of freedom for the original multinomial logistic model. (**2 points**).

6. Using a subset of the `Boston` data from class, we model median home value (`medv`) as a nonlinear function of the percentage of low-socieconomic-status residents in each census tract (`lstat`). We use cubic polynomial regression, with an intercept and regressors `lstat`, `lstat`^{2}, and `lstat`^{3}. The fitted coefficients for the different basis elements are as follows:

Basis element	Estimate	Std. Error
(Intercept)	48.93	1.53
`lstat`	-3.733	0.349
`lstat`^{2}	0.1419	0.0226
`lstat`^{3}	-0.001915	0.000427

Residual standard error: 5.397 on 446 degrees of freedom

(a) Compute the fitted value for median home value for a census tract with 10 percent low-socioeconomic status residents. (**4 points**).

(b) Write down an expression giving an approximate 95% confidence interval for the fitted value in the previous part, as a function of $(X^T X)^{-1}$,where X is the design matrix for the polynomial basis (including the intercept term). Please substitute numbers for the other parts of the expression. (**6 points**).

(c) Suppose now we fit a cubic regression spline instead of a polynomial, introducing one knot at the median value of `lstat`. How should we change the design matrix to fit this model instead? How many degrees of freedom will the model now use? (**4 points**).

7. Answer TRUE or FALSE to the following statements **and justify your answer**.

(a) The leverage for the ith subject measures how far the ith subject is from the rest of the subjects in terms of the explanatory variable values. (**2 points**).

(b) The MLE of β in a logistic regression model can always be computed in closed form. (**2 points**).

(c) Transformations of the explanatory variables can potentially improve the fit of the model in logistic regression. (**2 points**).

(d) In a model with p continuous explanatory variables, Mallow's C_p must be $\geq p + 1$. (**2 points**).

(e) R^2 is an effective model selection criterion for deciding the best size for a linear model. (**2 points**).

(f) In simple regression, rejecting the null hypothesis that $\beta = 0$ is equivalent to concluding that x has a causal effect on y. (**2 points**).