Practice Questions for Final Exam

Statistics 151a, Fall 2016

04 December, 2016

1. Last year 80 students took this particular course at Berkeley of whom 20 were freshmen, 20 were sophomores, 20 juniors and 20 seniors. In R, I have saved the scores (out of a maximum 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] 58.53768
> sd(g1)
[1] 5.024681
> mean(g2)
[1] 64.72989
> sd(g2)
[1] 4.43851
> mean(g3)
[1] 64.06235
> sd(g3)
[1] 5.264511
> mean(g4)
[1] 66.27922
> sd(g4)
[1] 4.192543
```

Let $y_1, \ldots, y_n$ (for $n = 80$) denote the scores of the students. The instructor assumes the model

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

where (i) $\beta_0, \beta_1, \beta_2, \beta_3$ and β_4 are unknown parameters, (ii) x_{ij} takes the value 1 if the ith student is in year j and 0 otherwise, and (iii) $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables with unknown σ^2.

(a) Is β_2 estimable in this model? What about $\beta_2 + \beta_3 - 2\beta_4$? What about $\beta_2 + \beta_4$? Give reasons in each case. (3 points)

(b) What is a good estimate of σ and why? (4 points)

(c) Calculate the value of the F-statistic for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$ and indicate its distribution under the null hypothesis. (6 + 2 points)

(d) Mike is currently a sophomore who is taking this year's version of this course. Give a 95% prediction interval for Mike's score in this class. (4 points)

2. Consider the usual regression data with response values $y_1, \ldots, y_n$ and explanatory variable values $x_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, p$. The response vector is Y and the matrix of explanatory variables is X. Let me denote by M_1 the usual linear model for Y based on X.

The residuals in M_1 will be denoted by $\hat{e}_1, \ldots, \hat{e}_n$. Also the leverages in M_1 are $h_{11}, \ldots, h_{nn}$ and the Jackknife residuals (also called standardized predicted residuals) are $t_1, \ldots, t_n$.

I have a suspicion that the kth observation is an outlier. To formally check this, I decide to fit the following linear model to the data:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \gamma D_i + e_i$$

where the additional explanatory variable D_i takes the value 1 when $i = k$ and the value 0 for all other values of i. This model will be denoted by M_2. Note that it has one additional explanatory variable compared to M_1.

(a) Is it reasonable to test the hypothesis that the kth observation is an outlier in the model M_1 by testing the hypothesis $H_0 : \gamma = 0$ against $H_1 : \gamma \neq 0$ in the model M_2? Why or why not? **(2 points)**.

(b) Express the least squares estimate of γ in M_2 in terms of $\hat{e}_k$ and h_{kk}. **(5 points)**.

(c) We learned in class that the Jackknife residual t_k can be used to construct a test for testing the hypothesis that the kth observation is an outlier in the model M_1. Show that this test is equivalent to the t-test for testing $H_0 : \gamma = 0$ against $H_1 : \gamma \neq 0$ in the model M_2. **(5 points)**.

3. Consider the bodyfat dataset used extensively in class. Consider the following R code and R output:

```
> body = read.delim("bodyfat_corrected.txt", header = TRUE, sep = "")
> lmod = lm(BODYFAT ~ AGE + WEIGHT + HEIGHT + KNEE + BICEPS + WRIST, data = body)
> summary(lmod)

Call:
lm(formula = BODYFAT ~ AGE + WEIGHT + HEIGHT + KNEE + BICEPS +
    WRIST, data = body)

Residuals:
    Min      1Q  Median      3Q     Max
-21.965  -3.585  -0.189   3.712  13.909

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 48.50000   11.89242   4.078 6.14e-05 ***
AGE          0.22802    0.03058   7.457 1.53e-12 ***
WEIGHT       0.25427    0.02962   8.586 1.06e-15 ***
HEIGHT      -0.46670    0.10567  -4.416 1.51e-05 ***
KNEE        -0.05529    0.28114  XXXXXX    0.844
BICEPS       0.17002    0.19733   0.862    0.390
WRIST       -3.05511    0.60752  -5.029 9.54e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: XXXXX on 245 degrees of freedom
Multiple R-squared:  XXXXXX,Adjusted R-squared:  XXXXXX
F-statistic: XXXXX on 6 and 245 DF,  p-value: < 2.2e-16

> library(leaps)
> vs <-regsubsets(BODYFAT~AGE + WEIGHT + HEIGHT + KNEE + BICEPS + WRIST,body)
> rs <- summary(vs)
> rs$which
  (Intercept)   AGE  WEIGHT HEIGHT KNEE BICEPS WRIST
1        TRUE FALSE    TRUE  FALSE FALSE  FALSE XXXXX
```

```
2       XXXXX  TRUE   XXXXX  FALSE  FALSE  FALSE  FALSE
3       TRUE   TRUE   TRUE   FALSE  FALSE  FALSE  TRUE
4       TRUE   TRUE   TRUE   XXXXX  FALSE  FALSE  TRUE
5       TRUE   TRUE   TRUE   TRUE   XXXXX  TRUE   TRUE
6       TRUE   TRUE   TRUE   TRUE   TRUE   TRUE   TRUE
> rs$cp
[1] XXXXXX  58.516859  22.524842 3.788903   5.038676   7.000000
> rs$adjr2
[1] 0.3725513 0.4603244 XXXXXXX 0.5608771 0.5604378 0.5587133
> rs$rss
[1] 10985.974  XXXXXX  8237.182  7596.320  7573.134  7571.938
```

(a) There are five missing values (indicated by XXXXX) in the *summary(lmod)* output above. Fill them with proper reasoning. **(5 points)**.

(b) There are five missing values (indicated by XXXXX) in the *rs$which* output above. Fill them with proper reasoning. **(5 points)**.

(c) There are one missing value each in the outputs of *rs$cp*, *rs$adjr2* and *rs$rss*. Fill them with proper reasoning **(6 points)**.

(d) Based on the output above, which model would you use for regressing BODYFAT based on the explanatory variables and why? **(2 points)**.

(e) Calculate the F-statistic and indicate its null distribution for testing model 1 against model 2 where **(4 points)**

```
Model 1: BODYFAT ~ AGE + WEIGHT + WRIST
Model 2: BODYFAT ~ AGE + WEIGHT + HEIGHT + KNEE + BICEPS + WRIST.
```

4. Consider the email spam data that we used in class. The data consists of 4601 emails of which 1813 emails were identified as spam. The response variable, named *yesno*, takes the value *y* if the email is spam and *n* otherwise. The explanatory variables include *crl.tot* (total length of words in capitals), *dollar* (number of occurences of the $ symbol), *bang* (number of occurences of the ! symbol), *money* (number of occurences of the word *money*), *n000* (number of occurences of the string *000*) and *make* (number of occurences of the word *make*). As in class, I fitted a logistic regression model to the data via

```
s = 0.001
M1 <- glm(yesno~ log(crl.tot) + log(dollar+s) + log(bang+s)
                 +log(money+s) + log(n000+s) + log(make+s),
                 family=binomial, data=spam)

summary(M1)
```

This gave me the following output:

```
Call:
glm(formula = yesno ~ log(crl.tot) + log(dollar + s) + log(bang +
    s) + log(money + s) + log(n000 + s) + log(make + s), family = binomial,
    data = spam)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1657  -0.4367  -0.2863   0.3609   2.7152

Coefficients:
             Estimate Std. Error z value
(Intercept)   4.11947    0.36342 XXXXXX
```

```
log(crl.tot)     0.30228    0.03693   8.185
log(dollar + s)  0.32586    0.02365  13.777
log(bang + s)    0.40984    0.01597  25.661
log(money + s)   XXXXXX     0.02800  12.345
log(n000 + s)    0.18947    0.02931   6.463
log(make + s)   -0.11418    0.02206  -5.177
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: XXXXXX  on XXXX  degrees of freedom
Residual deviance: 3245.1  on XXXX  degrees of freedom
AIC: XXXXXX

Number of Fisher Scoring iterations: 6
```

(a) Fill the six missing values in the above output giving appropriate reason. **(6 points)**

(b) Suppose a new email comes in for which

crl.tot	dollar	bang	money	n000	make
157	0.868	2.894	0	0	0

According to the above logistic regression model, what is the predicted probability that this email is spam? **(2 points)**.

(c) It may be noted that in the model $M1$, I took logarithms of the explanatory variables. I decided to fit another logistic regression model without taking logarithms of the explanatory variables:

```
M2 = glm(yesno~ crl.tot + dollar+ bang +money+n000 +make, family=binomial, data=spam)
```

The residual deviance for this model turned out to be 4058.8. On the basis of this, which of the two models $M1$ and $M2$ would you use and why? **(2 points)**.

(d) For each threshold/cut-off value in the set $\{0.05, 0.1, \ldots, 0.9, 0.95\}$, I calculated the precision and recall of both models $M1$ and $M2$. This resulted in the ROC curve shown in Figure 1. Which of the two models $M1$ and $M2$ would you prefer based on this ROC curve and why? **(3 points)**.

5. Consider the email spam data spam7 that we used in class. The data consists of 4601 emails 1813 of which were identified as spam. The explanatory variables are *crl.tot*, *dollar*, *bang*, *money*, *n000* and *make*. The response variable takes the value y if the email is spam and n otherwise. I fit a classification tree to the dataset using the following R code:

```
library(DAAG)
data(spam7)
sprt = rpart(yesno ~ crl.tot + dollar + bang + money + n000 + make,
method = "class", data = spam7)
```

This gave me the following output:

```
> sprt
n= 4601

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 4601 1813 n (0.6059552 0.3940448)
   2) dollar< 0.0555 3471  816 n (0.7649092 0.2350908)
```

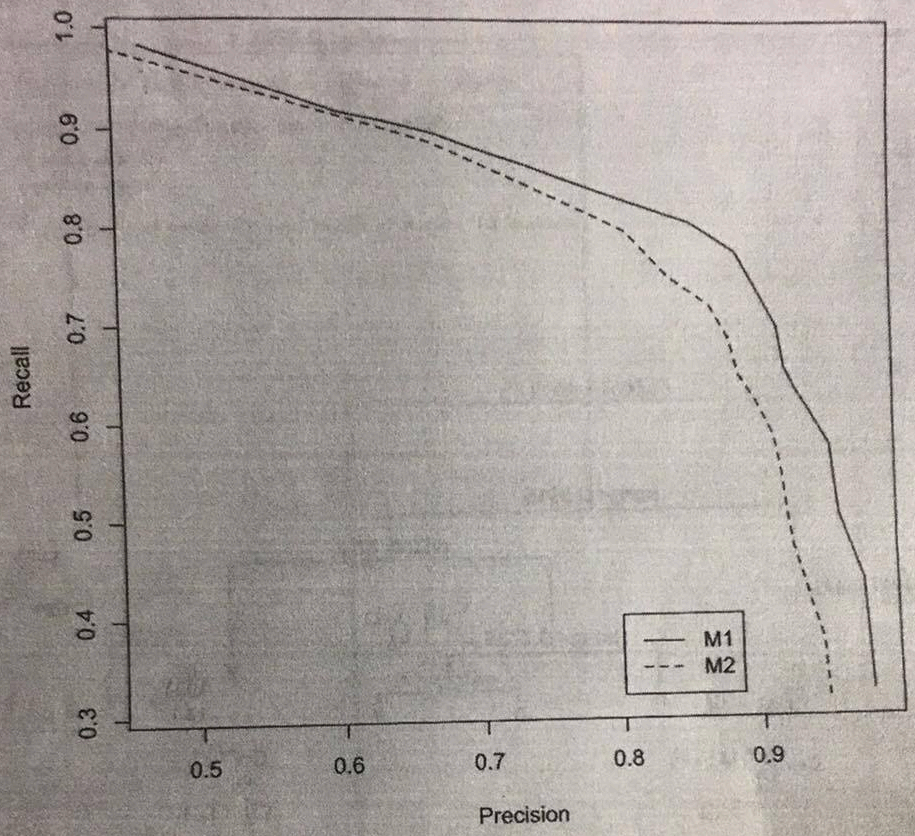

Figure 1: ROC curve.

```
4) bang< 0.0915 2420  XXX XX (0.8983471 0.1016529) *
5) bang>=0.0915 1051  481 y (0.4576594 0.5423406)
  10) crl.tot< 85.5 535  175 n (0.6728972 0.3271028)
    20) bang< 0.7735 XXX  106 n (0.7464115 0.2535885) *
    21) bang>=0.7735 117   48 y (0.4102564 0.5897436)
      42) crl.tot< 17 43   12 n (0.7209302 0.2790698) *
      43) crl.tot>=17 74   17 y (XXXXXXXX  XXXXXXX) *
  11) crl.tot>=85.5 516  121 y (0.2344961 0.7655039) *
3) dollar>=0.0555 1130  133 y (0.1176991 0.8823009) *
```

I then tried to plot this tree via

```
plot(sprt)
text(sprt)
```

which gave me the plot in Figure 2.

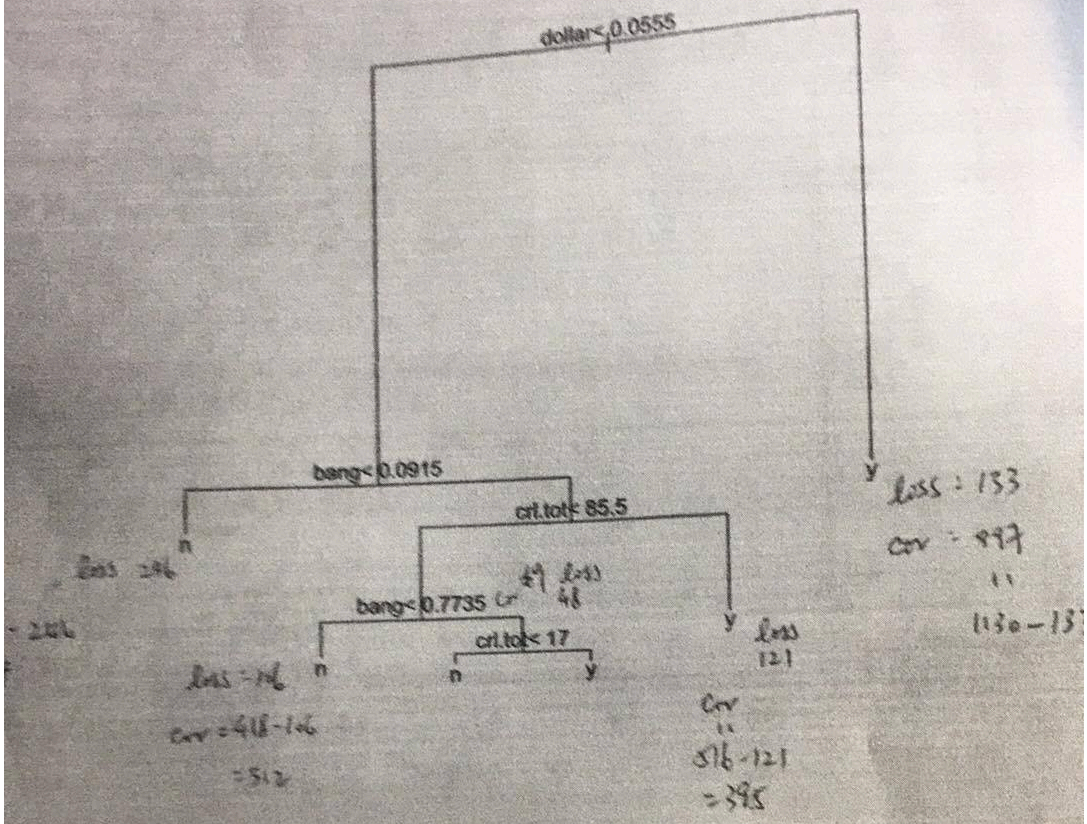

Figure 2: The tree *sprt*

(a) There are five missing values (indicated by the X symbol) in the above output for *sprt*. Fill them giving reasons (**5 points**).

(b) What is the RSS for *sprt*? (**3 points**)

(c) Consider the following R code and output:

```
> printcp(sprt)

Classification tree:
rpart(formula = yesno ~ crl.tot + dollar + bang + money + n000 +
    make, data = spam, method = "class")
Root node error: 1813/4601 = 0.39404
n= 4601

        CP nsplit rel error  xerror     xstd
1 0.476558      0   1.00000 1.00000 0.018282
```

```
2 0.075565      1   0.52344 0.54661 0.015380
3 0.011583      3   0.37231 0.38555 0.013429
4 0.010480      4   0.36073 0.37893 0.013334
5 0.010000      5   0.35025 0.37452 0.013270
```

Based on the above, I decided to prune *sprt* with $cp = 0.011583$. Is this reasonable? **(2 points).**

(d) The pruned tree is plotted in Figure 3 as follows:

```
p.sprt = prune(sprt, cp = 0.011583)
plot(p.sprt)
text(p.sprt)
```

Calculate the precision and recall of *p.sprt*. **(4 points).**

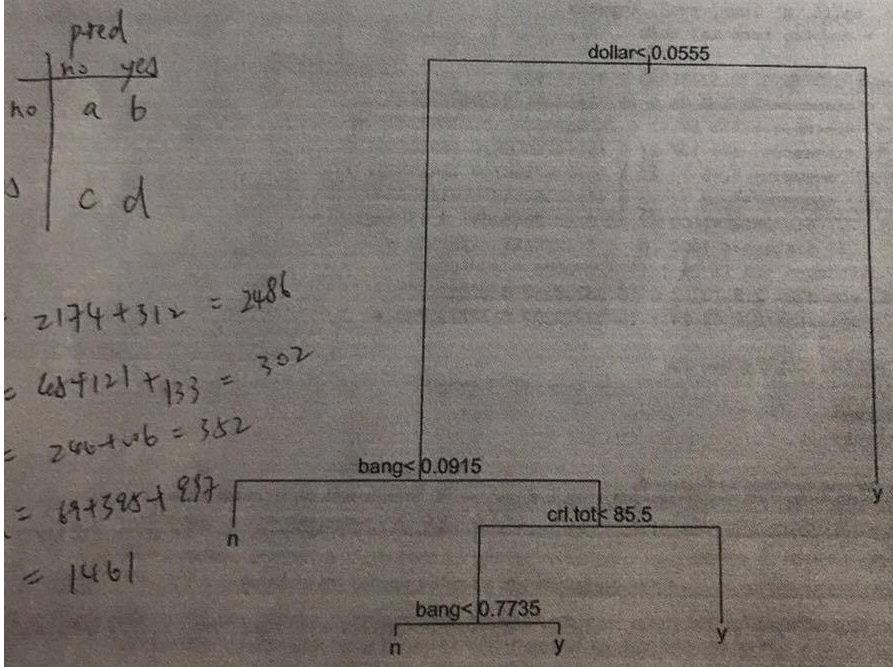

Figure 3: The Pruned Tree *p.sprt*

6. Consider the frogs dataset that we used in class. To describe the data briefly, 212 sites of the Snowy Mountain area of New South Wales, Australia were surveyed for the species of the Southern Corroboree frog. The response variable, named *pres.abs*, takes the value 1 if frogs of this species were found at

the site and 0 otherwise. The explanatory variables include *altitude*, *distance*, *NoOfPools*, *NoOfSites*, *avrain*, *meanmin* and *meanmax*. The dataset contains 212 observations and the response variable equals one for 79 observations and equals 0 for the rest. I fit a classification tree to the dataset using the following R code:

```
library(DAAG)
data(frogs)
ctree = rpart(pres.abs ~ altitude + distance + NoOfPools + NoOfSites +
avrain + meanmin + meanmax, method = "class", data = frogs)
```

This gave me the following output:

```
> ctree
n= 212

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 212 79 0 (0.62735849 0.37264151)
   2) distance>=625 137 28 0 (0.79562044 0.20437956)
     4) distance>=3375 30  1 0 (0.96666667 0.03333333) *
     5) distance< 3375 107 27 0 (0.74766355 0.25233645)
      10) meanmin< 3.15 76 XX X (0.81578947 0.18421053) *
      11) meanmin>=3.15 31 13 0 (0.58064516 0.41935484)
        22) distance>=1600 XX  2 0 (0.86666667 0.13333333) *
        23) distance< 1600 16  5 1 (XXXXXX  XXXXXX) *
   3) distance< 625 75 24 1 (0.32000000 0.68000000)
     6) meanmin< 2.9 12  2 0 (0.83333333 0.16666667) *
     7) meanmin>=2.9 63 14 1 (0.22222222 0.77777778) *
```

I then tried to plot this tree via

```
plot(ctree)
text(ctree)
```

which gave me the plot in Figure 4.

(a) There are five missing values (indicated by the X symbol) in the above output for *ctree*. Fill them giving reasons (**5 points**).

(b) What is the RSS for ctree? (**3 points**)

(c) For what values of α, does the inequality $C_\alpha(\text{ctree}) \geq C_\alpha(\text{root tree})$ hold? Here $C_\alpha(T)$ is defined as $RSS(T) + \alpha|T|TSS$. (**2 points**).

(d) What are the precision and recall for this classification tree? (**4 points**)

(e) Suppose I decide to use the variable log(distance) as opposed to distance. In other words, I construct the tree via

```
logctree = rpart(pres.abs ~ altitude + log(distance) + NoOfPools + NoOfSites +
avrain + meanmin + meanmax, method = "class", data = frogs)
```

Manually draw and label this tree. Give reasons when making claims. (**3 points**)

7. Determine whether each of the following statements is true or false. Provide reasons in each case. (**14 points**)

(a) For estimating β in the model $Y \sim N_n(X\beta, W)$, the estimator $(X^TWX)^{-1}(X^TWY)$ is preferable to the estimator $(X^TX)^{-1}X^TY$.

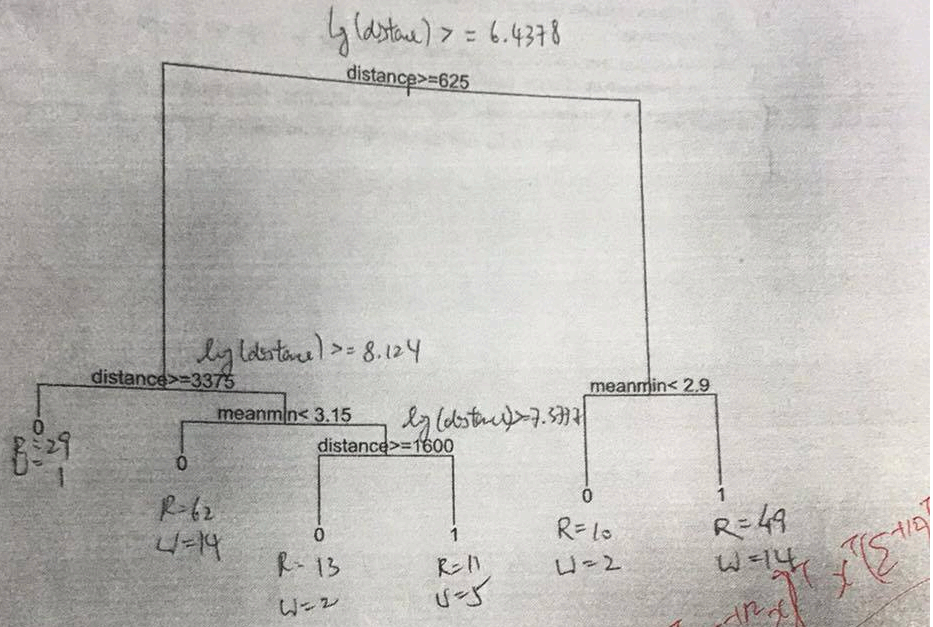

Figure 4: The tree *ctree*

(b) For estimating β in the model $Y \sim N_n(X\beta, W)$, the estimator $(X^TW^{-1}X)^{-1}(X^TW^{-1}Y)$ is preferable to the estimator $(X^TX)^{-1}X^TY$.

(c) The residual bootstrap algorithm for giving confidence intervals for $\beta_2 - \beta_3$ in the linear model $Y = X\beta + e$ requires no assumptions on the distribution $e_1, \ldots, e_n$ at all to work.

(d) When the sample size n is large and when there are no points with high leverages, the vector of residuals $\hat{e}$ can be seen as a proxy to the unknown vector of errors e.

(e) 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.

(f) Partial regression and partial residual plots can be used to check the assumption of linearity in the linear regression model.

(g) The sum of the fitted probabilities in a logistic regression model equals the sum of the response values.

(h) The sum of the fitted probabilities in a linear regression model equals the sum of the response values.

(i) The MLE of β in a logistic regression model can always be computed in closed form.

(j) Newton's method for computing the MLE in a logistic regression model is equivalent to iteratively reweighted least squares.

T (k) Transformations of the explanatory variables can potentially improve the fit of the model in logistic regression.

T (l) When multiple outliers are present in a dataset, it might not be possible to detect them by looking at leverages or Cook's distance.

F (m) When the X matrix in linear regression is singular, the adjusted R^2 is not well-defined.

F (n) For algorithms designed for solving classification problems, precision is more important than recall.

r_i^2 h_i $\frac{r_i^2}{p+1}$ $\frac{h_i}{1-h_i}$

X invertible (X is not a full-rar

$$1 - \frac{RSS/n-p(m)-1}{TSS/n-1}$$