Final

Statistics 151, Fall 2013

18 December, 2013

1. Consider the body fat dataset that we used extensively in class. I want to fit the model:

   $$BODYFAT = \beta_0 + \beta_1 AGE + \beta_2 WEIGHT + \beta_3 HEIGHT + \beta_4 (WEIGHT + 3 * HEIGHT) + \beta_5 WRIST + e$$

   which I accomplish by the following R code resulting in the output given below:

```
> model = lm(BODYFAT ~ AGE + WEIGHT + HEIGHT + I(WEIGHT + 3*HEIGHT) + WRIST, data = body)
> summary(model)

Call:
lm(formula = BODYFAT ~ AGE + WEIGHT + HEIGHT + I(WEIGHT + 3 *
    HEIGHT) + WRIST, data = body)

Residuals:
     Min       1Q   Median       3Q      Max
-20.5918  -3.3673  -0.0016   3.4240  12.8823

Coefficients: (1 not defined because of singularities)
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            47.21461    8.89363   5.309 2.46e-07 ***
AGE                     0.20629    0.02807   7.349 2.91e-12 ***
WEIGHT                  0.24341    0.01672  14.562  < 2e-16 ***
HEIGHT                 -0.44389    0.09706  -4.574 7.59e-06 ***
I(WEIGHT + 3 * HEIGHT)       NA         NA      NA       NA
WRIST                  -2.73998    0.55167  -4.967 1.27e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.142 on 247 degrees of freedom
Multiple R-squared: 0.5669,Adjusted R-squared: 0.5599
F-statistic: 80.82 on 4 and 247 DF,  p-value: < 2.2e-16
```

   (a) Why does R produce NAs in the output? (**2 points**)

   (b) The estimate for β_2 is apparently 0.24341. Does this make sense? Explain. (**3 points**)

   (c) I decide against including the variable $WEIGHT + 3 * HEIGHT$ in the model and just intend to fit

   `Model M: BODYFAT ~ AGE + WEIGHT + HEIGHT + WRIST`

   What is the RSS for this model? Why? (**2 points**)

   (d) The model M has too many parameters for my liking; so I decide to consider the following model:

   `Model m: BODYFAT ~ AGE + WEIGHT`

which gave me the following R output:

```
Call:
lm(formula = BODYFAT ˜ AGE + WEIGHT, data = body)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3171  -4.3293   0.2917   3.9898  18.5237

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -18.37392    2.57545  -7.134 1.06e-11 ***
AGE           0.18269    0.02853   6.403 7.54e-10 ***
WEIGHT        0.16271    0.01224  13.298  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.696 on 249 degrees of freedom
Multiple R-squared: 0.4642,Adjusted R-squared: 0.4599
F-statistic: 107.9 on 2 and 249 DF,  p-value: < 2.2e-16
```

Find the p-value for testing the model m against the model M. (**4 points**).

(e) Calculate the values of Mallows's C_p for m and M. Which of these two models would you prefer according to the C_p criterion? (**3 points**).

2. For the bodyfat dataset used in class, I ran a regression and obtained the following output:

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

Residuals:
    Min      1Q  Median      3Q     Max
-20.419  -3.295  -0.183   3.443  12.711

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.49674   11.02797   4.216 3.50e-05 ***
AGE          0.20927    0.02836   7.380 2.45e-12 ***
WEIGHT       0.23705    0.02746   XXXXX 7.79e-16 ***
HEIGHT      -0.43386    0.09799  -4.427 1.44e-05 ***
KNEE        -0.06969    0.26070  -0.267    0.789
BICEPS       0.14925    0.18299   0.816    0.416
WRIST       -2.80080    0.56336  -4.972 1.25e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 5.155 on 245 degrees of freedom
Multiple R-squared: 0.5682,Adjusted R-squared: XXXXX
F-statistic: XXXX on XX and XXX DF,  p-value: < 2.2e-16
```

(a) Fill the five missing values giving proper reasons (**5 points**).

(b) Look at the diagnostic plot in Figure 1. Based on this plot, is observation 42 an outlier? influential point? Explain. (**2 points**)

(c) The values of the variables for observation 42 are given by

BODYFAT	AGE	WEIGHT	HEIGHT	KNEE	BICEPS	WRIST
31.7	44	205	29.5	42.5	33.6	17.4

Is there anything unusual about this observation? (**1 point**).

(d) I decided to drop observation 42 and perform the regression again. Here is what I got:

```
Call:
lm(formula = BODYFAT ~ AGE + WEIGHT + HEIGHT + KNEE + BICEPS +
    WRIST, data = body[-42, ])

Residuals:
    Min      1Q  Median      3Q     Max
-23.370  -3.227   0.012   3.264  10.935

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.45236   12.35349   5.946 9.46e-09 ***
AGE          0.18009    0.02822   6.383 8.69e-10 ***
WEIGHT       0.24990    0.02670   9.361  < 2e-16 ***
HEIGHT      -0.94291    0.15138  -6.229 2.04e-09 ***
KNEE         0.18832    0.25885   0.728    0.468
BICEPS       0.03573    0.17871   0.200    0.842
WRIST       -2.71247    0.54457  -4.981 1.20e-06 ***
---
```

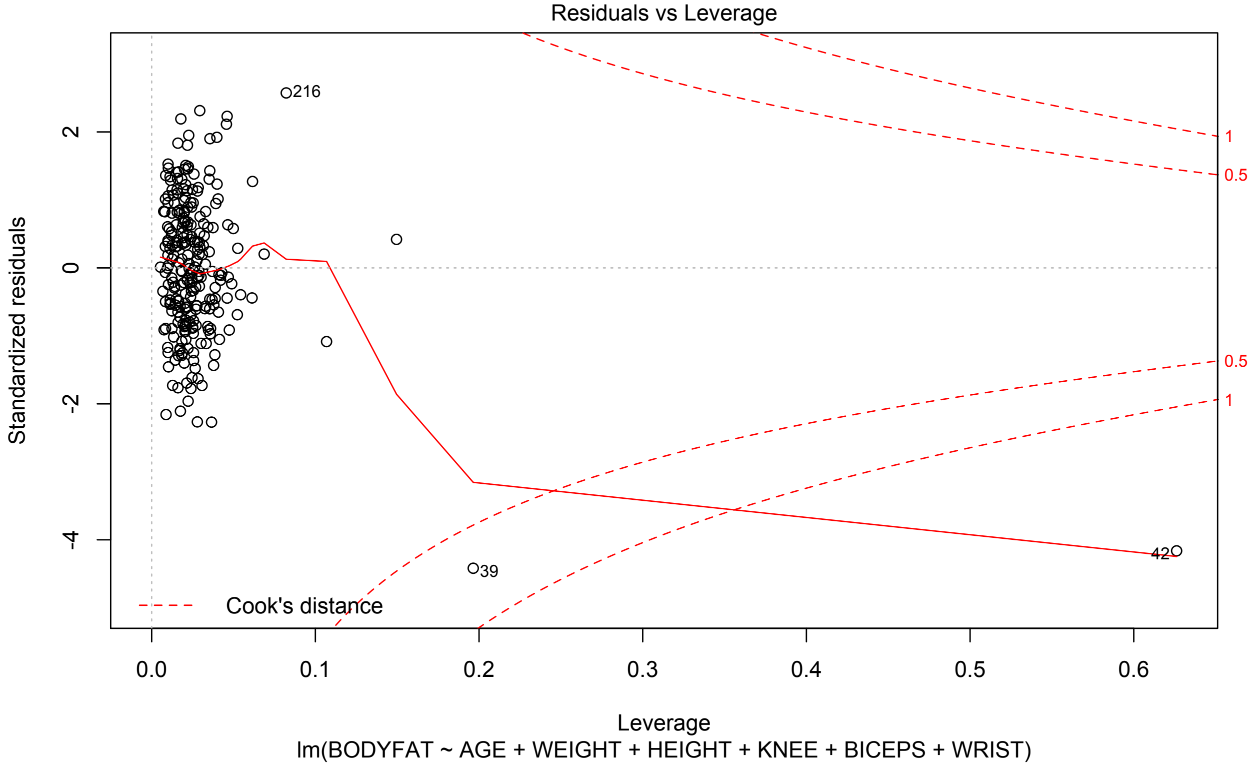

Figure 1: A regression diagnostic plot

.

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.98 on 244 degrees of freedom
Multiple R-squared: 0.5943,Adjusted R-squared: 0.5844
F-statistic: 59.58 on 6 and 244 DF,  p-value: < 2.2e-16
```

Based on the above two regression outputs, describe a test for assessing whether the observation 42 is an outlier (in the first regression) and calculate its p-value. (**6 points**)

3. 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 logistic regression model to the data via

```
frogs.glm <- glm(formula = pres.abs ~ log(distance) +
                  log(NoOfPools) + meanmin,
                  family = binomial, data = frogs)
summary(frogs.glm)
```

This gave me the following output:

```
Call:
glm(formula = pres.abs ~ log(distance) + log(NoOfPools) + meanmin,
    family = binomial, data = frogs)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9642  -0.7657  -0.4619   0.8728   2.3219

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.6864     XXXXX    0.313 0.754146
log(distance)   -0.9050     XXXXX   -4.349 1.37e-05 ***
log(NoOfPools)   0.5027     0.2004   2.509 0.012102 *
meanmin          1.1153     0.3131   3.562 0.000369 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: XXXXX  on XXX  degrees of freedom
Residual deviance: XXXXX  on XXX  degrees of freedom
AIC: 222.18

Number of Fisher Scoring iterations: 5
```

Also consider the following R code:

```
X = model.matrix(frogs.glm)
W = diag(frogs.glm$fitted.values*(1 - frogs.glm$fitted.values))
solve(t(X) %*% W %*% X)
```

which gave me the output

```
              (Intercept) log(distance) log(NoOfPools)     meanmin
(Intercept)      4.8038479  -0.363947754   -0.255928180 -0.49698440
log(distance)   -0.3639478   0.043313307    0.008053415  0.01562971
log(NoOfPools)  XXXXXXXXXX   0.008053415    0.040141698  0.02678507
meanmin         -0.4969844   0.015629708    0.026785069  XXXXXXXXX
```

(a) Fill the eight missing values in the above output giving appropriate reasons. (**8 points**)

(b) Suppose a new site is found where the values of the explanatory variables are

```
distance = 265      NoOfPools = 26       meanmin = 3.5
```

According to the logistic regression model, what is the predicted probability that Southern Corroboree frogs will be found at this site? (**2 points**).

(c) Suppose I add the variable *altitude* to the model. Would the residual deviance increase or decrease? Explain with reason. Would the null deviance increase or decrease? Explain with reason. (**4 points**).

4. Consider again *frogs* dataset. 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 2.

(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**)

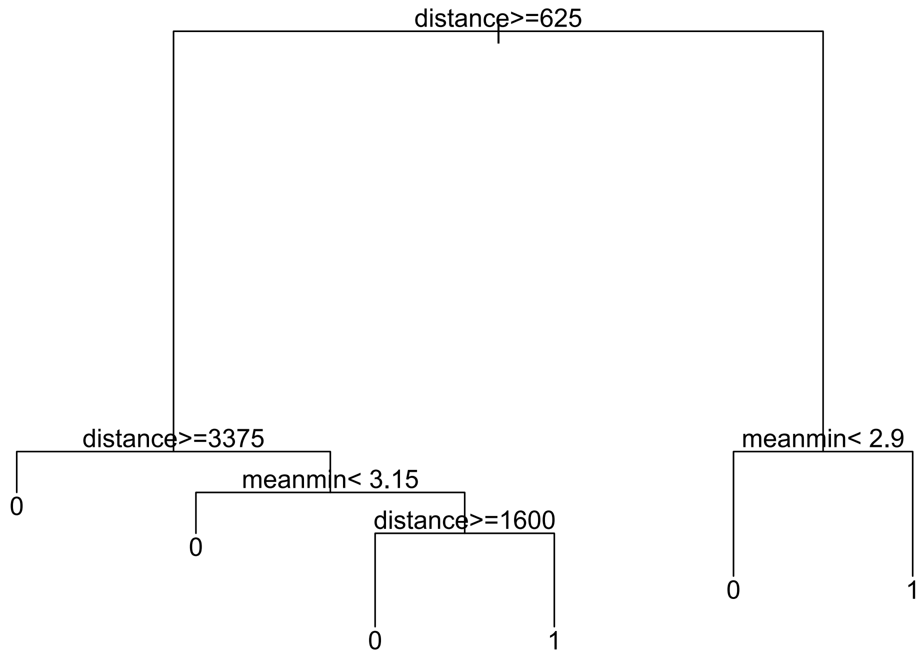

Figure 2: The tree *ctree*

.

5. Determine whether each of the following statements is true or false. Provide reasons in each case. (**10 points**)

   (a) The vector of fitted values is the Best Unbiased Linear Estimator of $X\beta$.

   (b) The vector of fitted values in a submodel is always worse as an estimator of $X\beta$ than the vector of fitted values of the full model.

   (c) The permutation test for testing $H_0 : \beta_1 = 0$ in the usual linear model requires the assumption of normality.

   (d) The mle in logistic regression is computed by a sequence of weighted least squares estimators.

   (e) $RSS_{[i]}$ and $\hat{e}_i$ are independent.

   (f) The sum of the squares of the residuals can be used as a model selection criterion in the linear model.

   (g) The sum of the squares of the predicted residuals can be used as a model selection criterion in the linear model.

   (h) In logistic regression, if the cut-off on the predicted probabilities is set too low or too high, the sum of the precision and recall will be large.

   (i) Multiple runs of the command printcp(tree) might produce different outputs.

   (j) The optimal tree for $cp = 0$ is always the root tree.