MSG500 Final 2010

December 16

The exam consist of 4 questions - all of them pertaining to the analysis of PCB levels in trout. It is probably easier to do the questions in order. Make sure to give detailed and specific answers. Avoid yes/no answers. You should also provide a motivation. Good Luck!

Question 1 a,b,c (25p)

Here I present some results from an analysis of PCB-levels in trout. The outcome is the PCB in the trout (on a log-scale), the predictors include the age of the fish, the weight and the length of the fish.
Using the figures and results below:
(a) Interpret the modeling results - what can you say about PCB levels in trout?
(b) Do you spot any problems with the fit? What, if anything, would you like to to change? What do you think the impact of this(these) change(s) would be?
(c) Do you spot any difficulties with modeling the PCB data? Comment on sample size, number of parameters, and predictor dependencies.

Below you can find scatter plots, basic diagnostic figures, changes in slopes and MSE, summary table of the fit, and a correlation matrix for the coefficient estimates ($V(\hat{\beta})$).

Summary of the model fit:

```
Call:
lm(formula = log(pcb) ~ length + weight + age, data = pcb)

Residuals:
      Min        1Q    Median        3Q       Max
-0.673818 -0.249682 -0.004904  0.385484  1.118320

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.2039275  0.7881886  -1.527   0.1423
length       0.0472198  0.0235607   2.004   0.0588 .
weight      -0.0002037  0.0003038  -0.671   0.5102
age          0.0920617  0.0670654   1.373   0.1850
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4833 on 20 degrees of freedom
Multiple R-squared: 0.7432,     Adjusted R-squared: 0.7047
F-statistic:  19.3 on 3 and 20 DF,  p-value: 4.037e-06
```

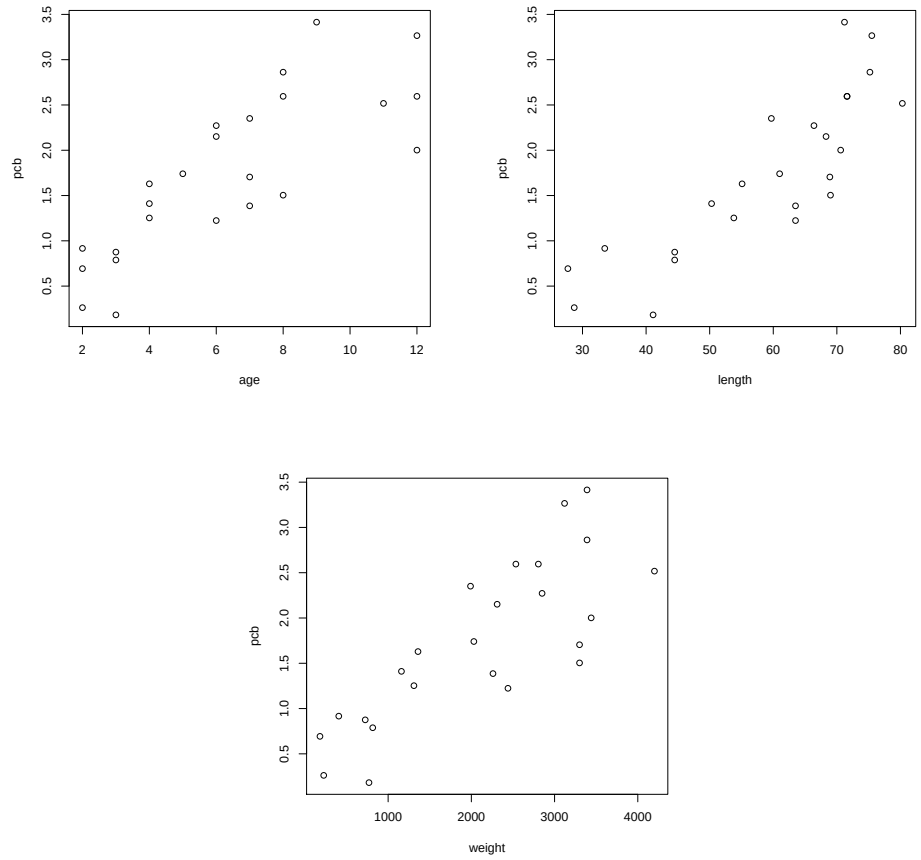

Figure 1: Scatterplots of log-pcb vs age and length and weight of the trout.

Correlation matrix for the coefficient estimates:

	(Intercept)	length	weight	age
(Intercept)	1.0000000	-0.9605785	0.7699946	0.1625124
length	-0.9605785	1.0000000	-0.8129483	-0.2718720
weight	0.7699946	-0.8129483	1.0000000	-0.2704618
age	0.1625124	-0.2718720	-0.2704618	1.0000000

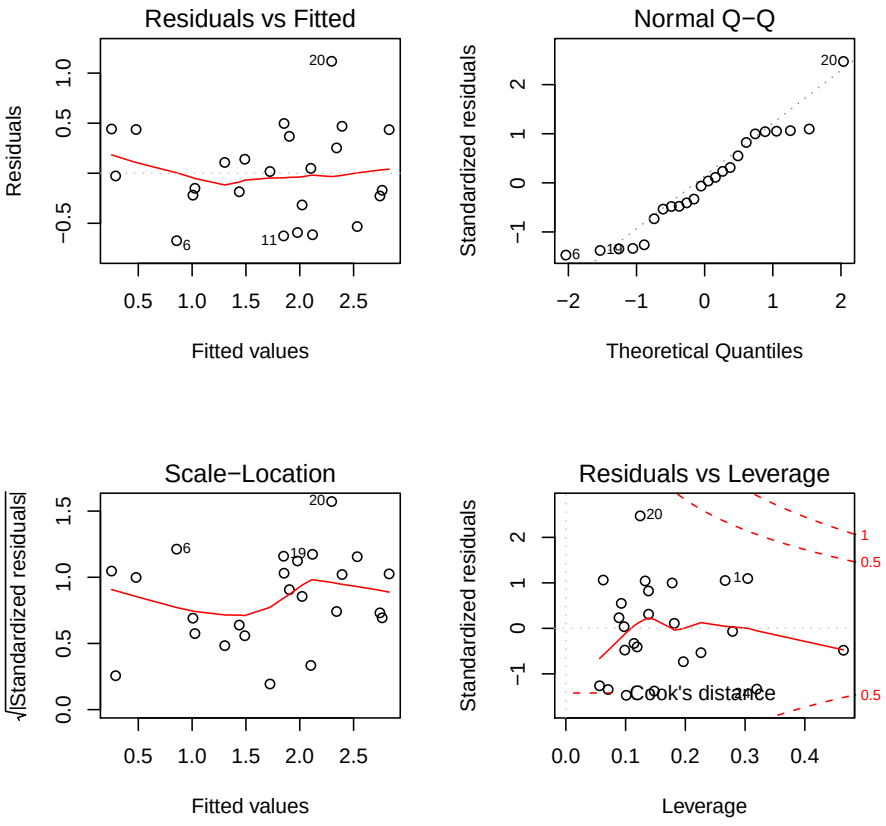

Figure 2: Diagnostic plots for the model including all 3 variables (age, length and weight)

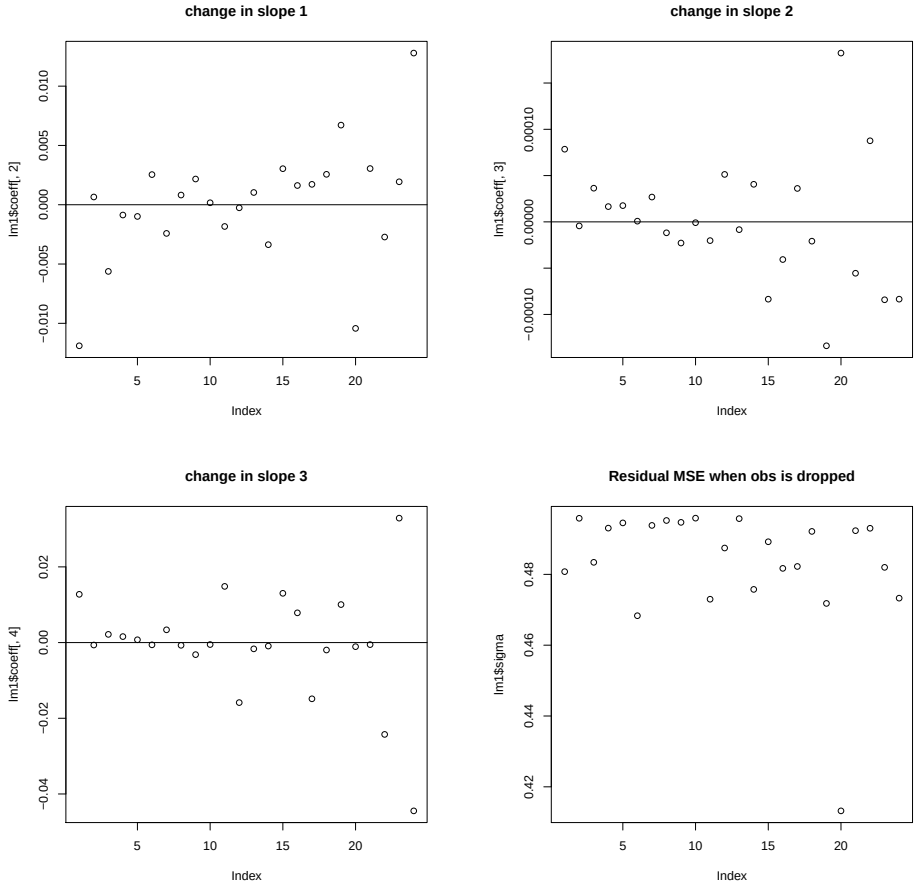

Figure 3: Impact of dropping observation i: Changes in slopes 1,2 and 3 (length, weight and age). Change in MSE.

Question 2 a,b,c,d,e,f (25p)

Continuing Question 1: I applied several model selection criteria to the PCB data. I kept 10% of the data for testing, and tried all-subset selection on the other 90% of data. The results are given here (Prediction error on the 10% test data, size of the selected model using Cp, AIC and BIC respectively, and which variables were selected).

```
"PE and size of selected models"
"PECP="  "0.0657" "size="  "2"
"PEAIC=" "0.1449" "size="  "3"
"PEBIC=" "0.0657" "size="  "2"
"CP model" "length"
"AIC model" "age"       "length"
"BIC model" "length"
```

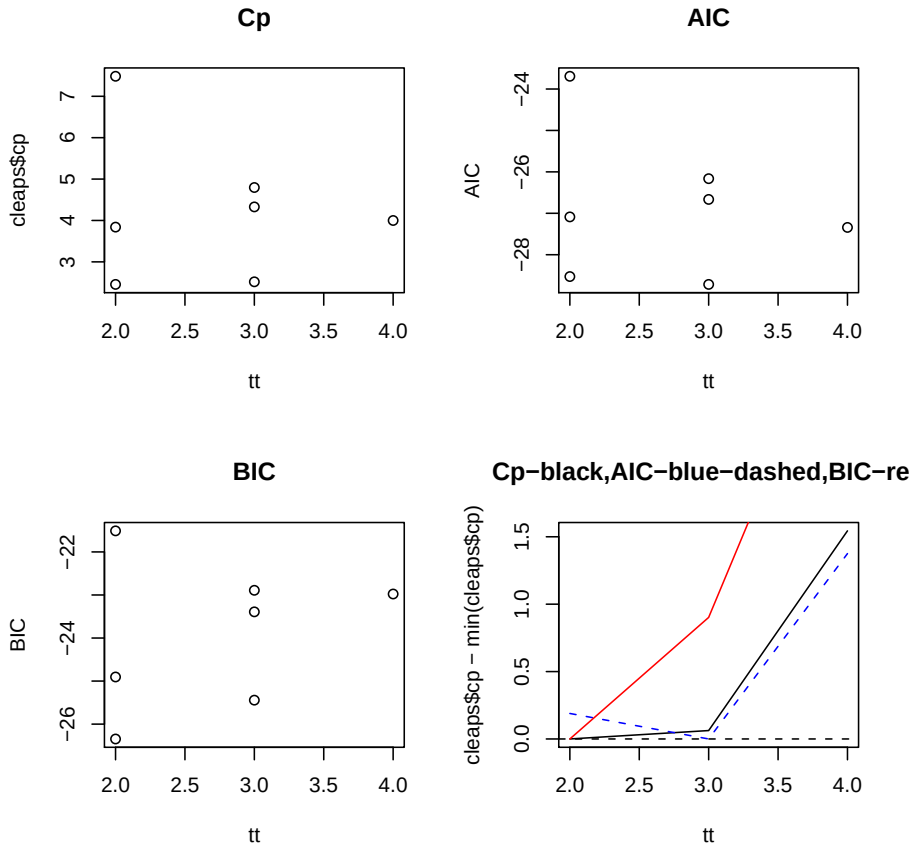

Figure 4: Cp, AIC and BIC values for all subset models (panels 1-3), the Cp, AIC and BIC curves (panel 4).

a)Please explain in detail what the panels in Figure 4 depict. What can you deduce from them? What can you conclude from the model selection procedure?

b)Comment on the size of the test data, the obtained prediction error estimates in the table, and how the first may affect the other.
c)Comment on the identified models, why and how they agree or disagree.

Part II I repeated the model selection procedure 100 times, each time leaving out a random 10% of the data for testing. The table below summarizes the outcome.

```
         modselcp modselaic modselbic
"age"    "35"     "48"      "6"
"length" "100"    "100"     "100"
"weight" "37"     "50"      "5"
"mean PE for cp, aic and bic"
"PEcp="    "0.2386"   " PEaic="  "0.2376"   " PEbicK=" "0.2064"
"mean model size for cp, aic and bic"
"sizecp="   "2.72"      " sizeaic=" "2.98"      " sizebic=" "2.11"
```

d) Please explain the table in detail. What can you tell about the models selected? Which predictor is the most important? How do the models selected by the three criteria (Cp, AIC and BIC) differ? Compare this result to the model summary in Question 1 - any surprises?

Part III I repeated the model selection 100 times, this time keeping 50% for training and model selection, and leaving out 50% of the data for testing. The table below summarizes the results.

```
          modselcp modselaic modselbic
"age"    "29"     "41"      "19"
"length" "79"     "83"      "69"
"weight" "24"     "35"      "20"
"mean PE for cp, aic and bic"
"PEcp="    "0.2929"   " PEaic="  "0.2921"   " PEbicK=" "0.2929"
"mean model size for cp, aic and bic"
"sizecp="   "2.32"      " sizeaic=" "2.59"      " sizebic=" "2.08"
```

e) Comment on the outcome of model selection in this case. What can you infer about the most important predictor here?
f) Compare the results in this table to the table above where 10% of the data was used for testing. Explain why the results differ in the way they do (pay attention to PE, model sizes and the selection frequencies).

Question 3 a,b,c (25p)

Continuing Questions 1 and 2: I also use CART to model the PCB data. I leave out 10% of the data for testing, train a CART model on the other 90%, run 10-fold CV on the training data to select the size of the tree, and then use this tree to predict the 10% test data. I repeat this twice. Below is the resulting 2 trees.
a) Explain both tree models. If I give you a trout aged 2 years, 65 cm long and weighing 720 grams - what do you predict its PCB level to be? (two predictions, one for each tree).

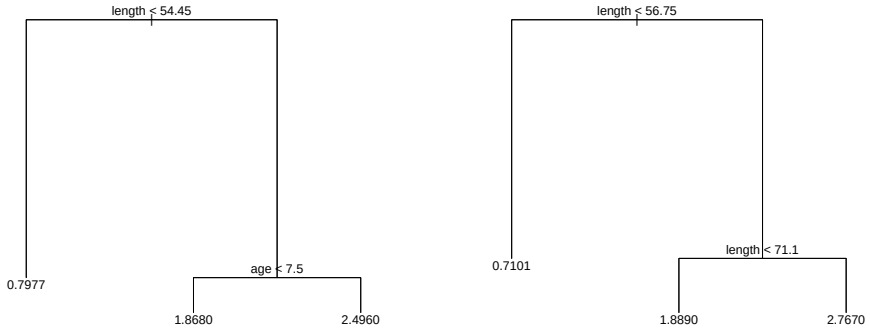

Figure 5: Two CART models obtain from two different 90% random subsets of data.

Part II I repeat the training and prediction process 100 times, each time leaving out a random 10% of data. The two tables below summarize (i) how many times each variable is picked to be in the tree (anywhere) and (ii) how many times the variable is picked to be the first split (top of the tree). The average size tree across 100 runs is 1.42, 58% of the trees contain only one split and the other 42% contain two splits. The mean prediction error is 0.2599.

```
     modtab
"age"    "0.37"
"length" "0.98"
"weight" "0.07"
    modfirst
"age"    "0.17"
"length" "0.78"
"weight" "0.05"
```

b) Comment on the results in the table, using the information I gave in the text as well.
c) How does this outcome compare with the outcome in Question 2? Which model is better, and in what sense: the linear model or CART? Which figure in the exam do you think best explains this and why?

Question 4 a,b,c (25p)

Continuing Questions 1-3: The PCB data contain several highly correlated variables (length and weight for example). In class we discussed regularized regression methods that could improve the fit in such situations. Here, we will revisit Principal Component (PC) Regression.
In PC regression we rotate the predictor variable coordinate system to form new variables: X becomes $\tilde{X}$. In the new coordinate system the $\tilde{x}$-variables are uncorrelated (the collinearity problem has been removed). The drawback is that each new variable, each $\tilde{x}$, is a linear combination of all the other variables.

I apply PC to the standardized x-variables in the PCB data (standardized = mean 0, standard deviation 1). I standardize the variables first so that we can compare the factor loadings directly (i.e. the weights given to each variable in the new $\tilde{x}$. The first PC component explains 94% of the variability in X. In the figure below I show you the factor loadings for each of the 3 principal components.

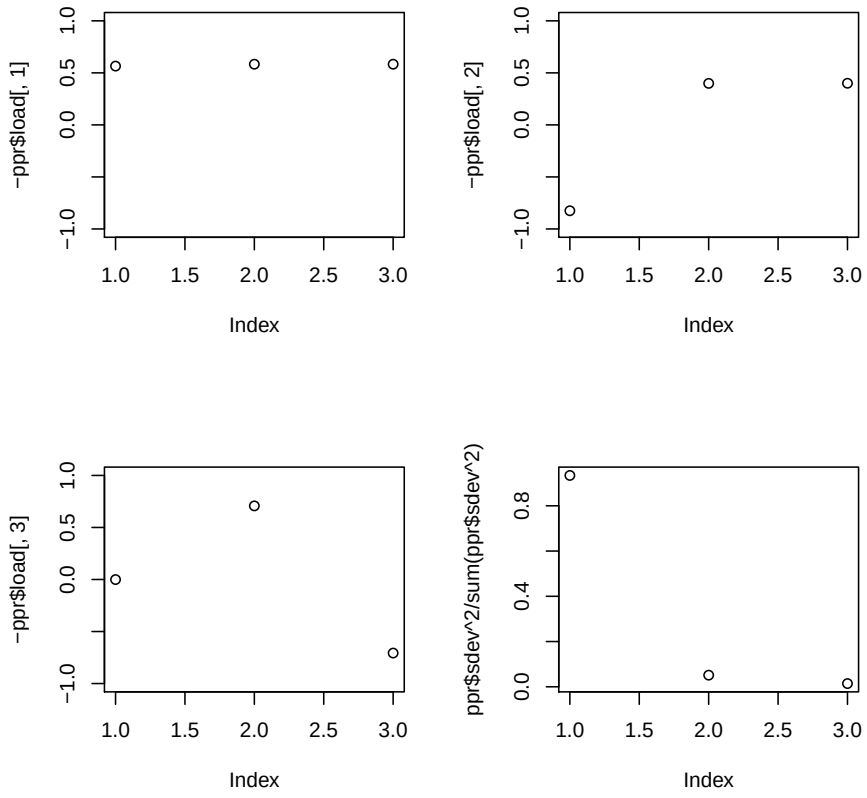

Figure 6: Panels 1-3: The PC loadings for components 1-3 (order of the variables is age, length and weight). Panel 4: the percent of variability explained by each of the 3 components.

In the figure you see that the first factor loading is near constant across the three variables. This means that the first component is essentially the mean of the 3 variables age, length and weight. The second component is essentially the difference between (length + weight)/2 and age (since the loadings are .4 for both length and weight and -.8 for age). The third component is the difference between length and weight (loadings .7 for length and -.7 for weight).
Below I summarize the linear model fit to the original data and then to the PC data.

```
Original fit:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.66014    0.08616  19.268 6.27e-14 ***
age          0.30460    0.18751   1.624    0.121
length       0.88834    0.31533   2.817    0.011 *
weight      -0.45523    0.31514  -1.445    0.165
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4132 on 19 degrees of freedom
Multiple R-squared: 0.7872,     Adjusted R-squared: 0.7536
F-statistic: 23.43 on 3 and 19 DF,  p-value: 1.339e-06

Principal component fit:
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.66014    0.08616  19.268  6.27e-14 ***
PC1          0.42482    0.05262  8.073   1.46e-07 ***
PC2         -0.07906    0.22441 -0.352   0.7285
PC3          0.94997    0.42518  2.234   0.0377 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4132 on 19 degrees of freedom
Multiple R-squared: 0.7872,     Adjusted R-squared: 0.7536
F-statistic: 23.43 on 3 and 19 DF,  p-value: 1.339e-06
```

a) Comment on the differences and similarities of the two models - explain what those differences and similarities mean.
b) Using my explanation of the loadings above, interpret the PC regression model (pay attention to the loadings, and the sign and magnitude of the estimated coefficients).
c) A stepwise model selection (backward model selection using AIC) leads to the following results: in the original fit, stepwise model selection does not lead to any variables being dropped; in the PC fit, stepwise model selection removes PC2 from the model and keeps PC1 and PC3. Can you explain why model selection leads to such different results depending if you use the original variables or the PCs?