MSG500 Solutions

Question 1 (30p)

Let's say you were given the following data set to analyze; number of observations n, the outcome y is continuous, and there are p predictor variables X.
You are asked to perform (i) model selection, and (ii) estimate the prediction error.
Say, for each of the following scenarios; **how** you would perform tasks (i) and (ii); **why** you recommend that particular approach for that particular scenario; and **concerns** you might have in each of these scenarios (e.g. depending on the characteristics of the data X or y).
a) $n = 250$, $p = 10$.
b) $n = 250$, $p = 100$.
c) $n = 30$, $p = 10$.

Brief answer:
a) This is near the ideal situation - roughly 25 observations per parameter to estimate. Model selection can be done with either of the criteria we discussed in class or by e.g. 10-fold cross-validation. To estimate the prediction error of a model you have to reserve part of the data set prior to modeling and model-selection. With 250 observations and 10 parameters, you can afford to hold out 25-50 observations for testing. You could also do this repeatedly on random splits of the data.
b) Here we have only 2.5 observations per parameter. In addition, this is a very high-dimensional problem. One concern you might have here is the suitability of the model across all the variables - there are a lot of scatter-plots and outlier diagnostics to check. In addition, you might be concerned about collinearities here - are some of the variables closely related? For model selection, you could try any of the criteria we have discussed in class. I would be hesitant to run cross-validation as the problem stands. In general, with such an unbalanced sample size to parameter setup, it is usually a good idea to try to simplify the problem a-priori. Can you group the variables and use just one variable per group? (Think about the car data- perhaps you can use just one size variable.) Also, you may want to consider a regularized model fitting strategy like PC regression, ridge regression or lasso. If you think that most of the variables are unrelated to the outcome you might want to approach the problem in a forward search manner rather than fitting the full model and try to select from it.
For prediction error estimation, unless you have first reduced the number of variables by either grouping them or using PCA, there is not much you can do. There is not much data to leave out for testing. You can try holding out 10% of the data for testing, but be aware that this could make the training more difficult and you only have a small sample to test on. If you have reduced the number of variables first, by PC or grouping, you are almost back in situation a).
c) This is situation b), but with even less flexibility. You have very few samples and comparatively many parameters. Like in b) you might look for redundancies, groups of variables, use regularized methods like PC or ridge or lasso. With so little data you cannot easily

afford to leave data out for testing.

Question 2 (30p)

Lets assume a student in a similar class to this one was asked to analyze the following data set: A health study conducted on n patients, where LDL (the bad cholesterol) is the variable of interest (the dependent variable). Information on p health related measures are also available (predictors). Examples of some of these p variables are; if the patient is on a low-fat diet or not; the age of the patient; number of hours per week that the patient exercises; body mass index (BMI), and bloodpressure, etc. The student chose to use a linear model to summarize LDL as a function of the predictor variables. Through residual analysis, the student found that the model was sufficient (no patterns in the residual plots), and that the normality assumption for the errors did not seem to be violated. The student that analyzed this data set put forth several claims or conclusions ((a) through (d) below). For each claim I want you to state whether you agree or disagree. I want you to motivate your choice. You may disagree on the basis of the statement being inaccurate - say in what sense. You may also state that the student is providing insufficient information to make the claim - if so, say what kind of additional information would be needed. If you agree, you should provide a similar motivation (why is the statement accurate, what kind of information is provided as supporting evidence).
(a) After fitting a regression model to the LDL data, I find that the lack-of-fit F-statistic is 10.2. I conclude that LDL is significantly related to some of the predictor variables in the data set.
(b) I set up a confidence interval for the slope coefficient related to bloodpressure using the quantiles of the t-distribution.. The confidence interval covers 0. I conclude that bloodpressure does not impact LDL.
(c) The p-value associated with the slope coefficient for BMI is 0.005. I conclude that BMI is significantly related to LDL.
(d) The R-squared of the full model is only 0.24. I conclude that a linear model using these health measures cannot be used to predict LDL.

Brief Answer:
a) Insufficient information. Since neither the number of variables or sample size is provided, we don't have the degrees-of-freedom for the F so we don't know what the critical values that denote significance are.
b) Almost correct. However, in testing we can only reject or fail to reject the null hypothesis. Here, we can only say that bloodpressure is not significantly related to LDL, but this could be due to lack of power (sample size too small). The correct phrase is that we cannot reject the hypothesis that bloodpressure is unrelated to LDL. Note also that "impact" sounds like a causative claim - whereas in regression we can only say something about associations. That is, it could be LDL that causus bloodpressure changes or vice versa.
c) Correct statement. But you guys also know to look for collinearities - are there variables in the data set related to BMI? If so, the p-value (and t-value it is based on) could be

misleading.
d) Incorrect. 24% of the variability of LDL is explained by the model. Yes, this is for the training data so perhaps we cannot *predict* at this level of precision. But the statement is much too strong. There is clearly information in the predictors about LDL, although a lot remains unknown as well.

Question 3 (30p)

The data set "baby" contains observations on 250 mothers and their newborns; bw: baby's weight at birth, to the nearest ounce; gd: gestation days (that is, total number of days of pregnancy); ma: mother's age in completed years; sm: indicator of whether the mother smoked (1) or not (0) during her pregnancy.
The main goal when analyzing this data set is to identify important predictors of low birth weight.

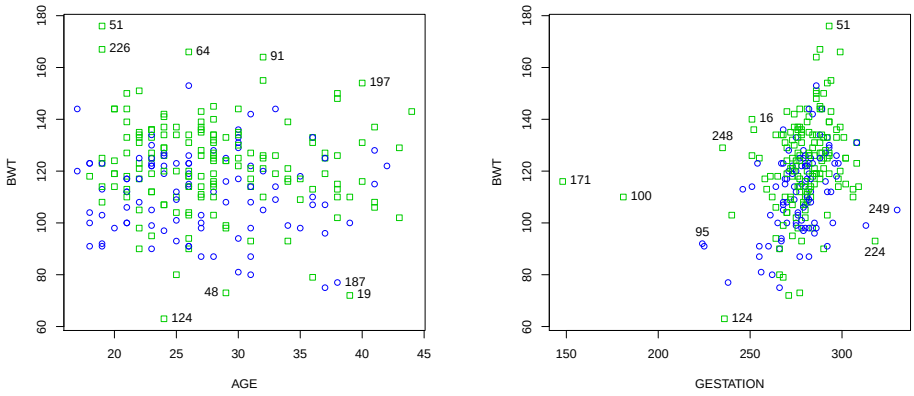

Figure 1: Scatterplots of birthweight vs age and gestation. Square symbols for non-smokers, circles for smokers. Observation numbers added to the plot.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.18697   17.13318   3.163  0.00176 **
smoke       -10.48311    2.15211  -4.871 1.99e-06 ***
age          -0.23831    0.16947  -1.406  0.16093
gestation     0.26897    0.05914   4.548 8.51e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 16.47 on 246 degrees of freedom
Multiple R-squared: 0.1632,     Adjusted R-squared: 0.153
```

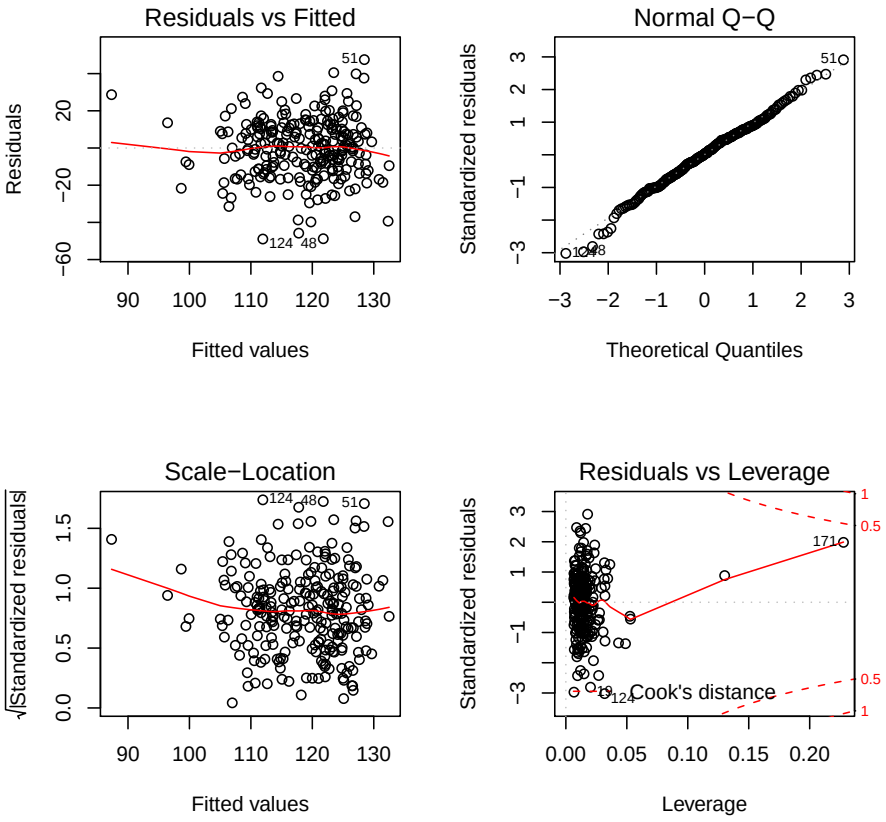

Figure 2: Diagnostic plots for the additive model

```
F-statistic: 15.99 on 3 and 246 DF,  p-value: 1.569e-09
```

From the scatterplots and diagnostic plots and output from an additive model fit, can you conclude that smoking has a negative impact on the birth weight? How does the mother's length of pregnancy impact the birth weight?
Are there any additional plots you feel are important to examine before drawing conclusions from the data?
Do you have any concerns regarding the fit?
Are there any unusual observations - if so, in what sense and how do you expect they impact the fit?

Brief Answer:
From the scatter plot of weight on age it is clear that the non-smokers lie above the smokers. That is, there seems to be an additive effect of smoking on birth-weight. The same is somewhat clear in the scatter plot of weight on gestation. A plot that would have helped illustrate this would be a box-plot of weight on smoking status. However, the scatter plots help show that whatever effect is due to smoking is not explained away by the other variables. Regarding birthweight and gestation, the answer is less clear. There are several extreme observations in terms of both response (birthweight) and predictor (gestation). In addition,

gestation is normally 40 weeks and we do not see much spread around this value. This makes it difficult to estimate the relationship between weight and gestation. If we trust the summary statistics in the table, the impact of gestation is significant and positive - that is, a longer pregnancy is associated with a higher birth weight.
Concerns: there are several outliers in this data sets - some with high leverage. Observations 171 and 100 are of particular concern. If we exclude them, gestation will be even more significant since these outliers will pull the regression slope down toward 0. The lower right panel in figure 2 shows clearly that we have a couple of high-leverage observations in our data. In addition, the R-squared is rather low for this data. We are probably missing some important predictors of low birth weight.

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      23.79979   27.84383   0.855 0.393523
smoke           -19.65874   40.88937  -0.481 0.631105
gestation         0.35250    0.09916   3.555 0.000454 ***
smoke:gestation   0.03613    0.14662   0.246 0.805556
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 16.41 on 244 degrees of freedom
Multiple R-squared: 0.1751,     Adjusted R-squared: 0.1649
F-statistic: 17.26 on 3 and 244 DF,  p-value: 3.375e-10
```

Here is the result from an interaction model. Please comment on this model output and compare with the additive model output. What can you conclude? To assist you I also provide the correlation matrix for the coefficient estimates (i.e., the scaled version of $V(\hat{\beta}) = \sigma^2(X'X)-1$).

```
                (Intercept)      smoke  gestation smoke:gestation
(Intercept)       1.0000000 -0.6809552 -0.9988566       0.6755093
smoke            -0.6809552  1.0000000  0.6801766      -0.9986086
gestation        -0.9988566  0.6801766  1.0000000      -0.6762826
smoke:gestation   0.6755093 -0.9986086 -0.6762826       1.0000000
```

(I also provide the correlation matrix for the additive model for comparison

```
            (Intercept)      smoke  gestation
(Intercept)   1.0000000 -0.1642298 -0.9978959
smoke        -0.1642298  1.0000000  0.1244658
gestation    -0.9978959  0.1244658  1.0000000
```

).

Brief Answer:
If we add an interaction term smoking-gestation to the additive model on smoking and gestation, the results indicate that only gestation is significant. The interaction term, i.e. testing if the impact of gestation on birth-weight is affected by smoking, is not significant. Note, this summary is based on the data including observations 100 and 171 (both non-smokers). While there is a small difference in slope of gestation for smokers and non-smokers, if observations 100 and 171 are dropped this interaction effect would be even less significant. Some concerns: the correlation matrices indicate that in the interaction model the estimates of effect of smoking on birth weight and the interaction effect are highly negatively correlated. What does that mean? That means that smoking and the interaction smoking:gestation are highly correlated variables. The same is true for the intercept and gestation, both in the

additive and interaction models. What is going on here? Well, gestation is in-fact near constant. In the additive model gestation thus almost plays the role of the intercept. The estimate of the gestation coefficient is therefore highly unstable. In the interaction model we have created another collinearity problem. Gestation is near-constant among the smokers and is thus highly correlated with the smoking variable itself. In the additive model we have thus essentially one clear-cut effect: smoking impacts birth weight. This estimate is not highly correlated with the other estimates. The gestation coefficient and the intercept are highly related, indicating that we have have to be careful about interpreting the magnitude (and perhaps even the sign) of this coefficient. A more in-depth study, using e.g. cross-validation, could help shed some light on whether gestation is predictive or an intercept is enough. (One could also perhaps make sure that gestation can only appear as a positive effect in the model to help regularize the fit).
Regarding the impact of smoking, in the interaction model we have lost the significance of both smoking and the interaction. We created a collinearity problem (remember the potatoes) and the result is that that nothing comes out significant. However, due to the excessive collinearity it might have been that either smoking or the interaction or both ended up significant. Comparing with the additive model structure, taking the scatter plots into account, it seems clear that we should not include the interaction term in the model.