MSG500 **Extra final**

February 18

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) $n = 30$, $p = 10$.

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.

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.

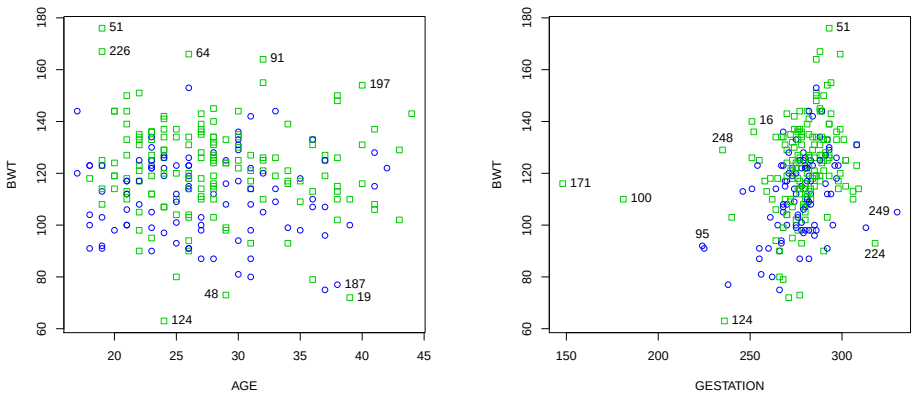

Figure 1: Scatterplots of birthweight vs age and gestation. Square symbols for non-smokers, circles for smokers. Observation numbers added to the plot.

```
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
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

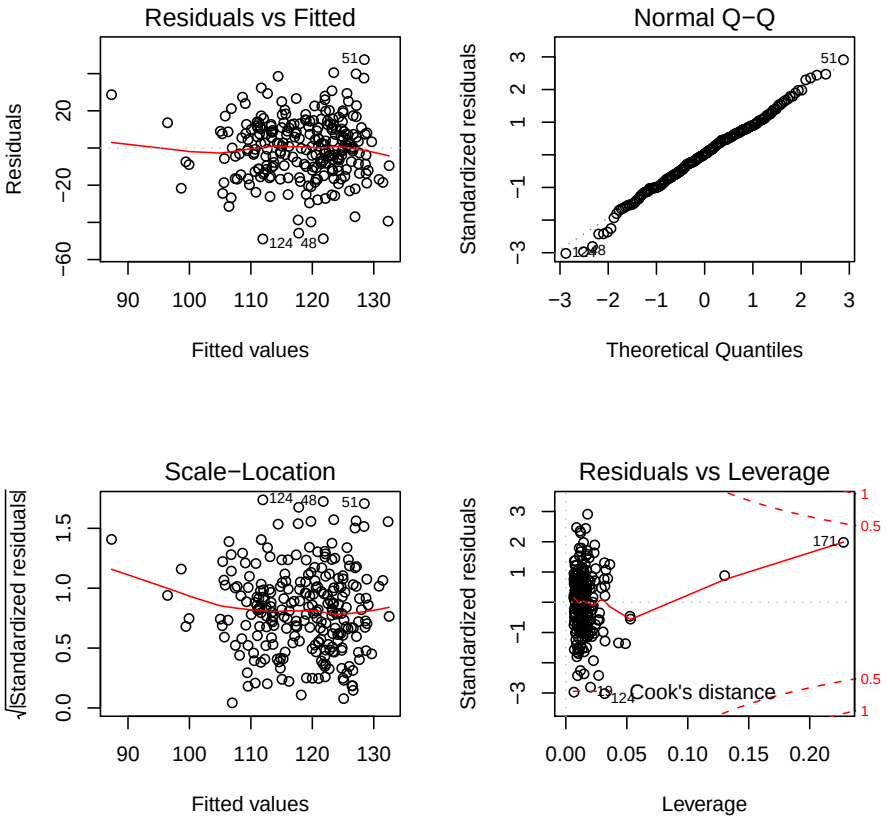

Figure 2: Diagnostic plots for the additive model

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?

```
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
```

).