PART I: Multiple Choice (5 Points Per Question). Unless otherwise instructed, choose the **best** answer.

1. A model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$ is fit to a dataset. Variables x_1 and x_2 have the relationship $x_1 = 3 + 2x_2$ with no error; that is, their correlation is 1. What will the added variable plot for variable x_1 look like?

   (a) A perfectly random scatter.

   (b) A straight horizontal line.

   (c) ***A straight vertical line. (Residuals from the model using x_2 as the response is on the x-axis.)

   (d) A straight line with slope = 2.

   (e) A straight line with slope = $\hat{\beta}_2$.

2. A nutritionist is interested in predicting the response variable insulin response from different types of breakfast cereals; volunteers are asked to eat cereal, and their insulin level is measured afterwards. The predictor variable of interest is x_1 = grams of fructose, but x_2 = grams of fat, x_3 = grams of glucose, and x_4 = grams of lactose are being controlled for. Which of the following is the correct method for testing whether fructose significantly affects insulin response? Assume model assumptions are met.

   (a) Simply test whether $\beta_1 = 0$ using the usual t-test.

   (b) First test whether $\beta_2 = \beta_3 = \beta_4 = 0$ using an F-test for model reduction. If we **fail to reject**, we can then test whether $\beta_1 = 0$ using a t-test.

   (c) First test whether $\beta_2 = \beta_3 = \beta_4 = 0$ using an F-test for model reduction. If we **reject**, we can then test whether $\beta_1 = 0$ using a t-test.

   (d) First test whether the overall model is significant: that is, test the null hypothesis that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. If we **fail to reject**, we can then test whether $\beta_1 = 0$ using a t-test.

   (e) ***First test whether the overall model is significant: that is, test the null hypothesis that $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. If we **reject**, we can then test whether $\beta_1 = 0$ using a t-test.

3. Researchers use a backward selection procedure based on p-values to choose a model, with significance level $\alpha = 0.10$. If they start with 60 variables randomly independently generated with **no population correlation with** y, how many are expected to remain in the model?

   (a) 0

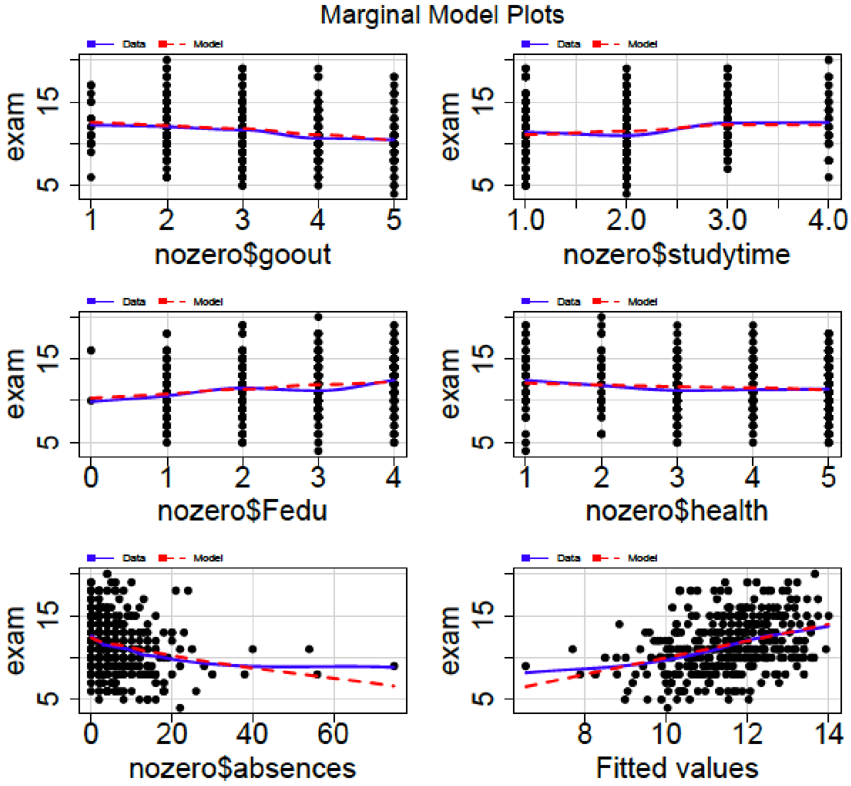
   (b) 2

   (c) 3

   (d) ***6

   (e) 36

4. Marginal model plots for the full model for students in Portugese schools from problem 6 below are shown here. What do these plots suggest?



(a) The linear model is as good as can be expected for random data.

(b) ***There seem to be just a few outliers not well predicted by the linear model.

(c) The first four variables are too discrete to be modeled well by the linear model.

(d) 'Go out' doesn't add anything to the model since its slope is negative.

(e) 'Absences' adds more to the model than 'study time' since the slope for 'absences' is steeper.

5. Which of the following is true about variable selection? AIC and BIC will choose the same model:

(a) ...among models with the same transformation of y, but different numbers of predictors.

(b) ...among models with the same transformations of the predictors, but different numbers of predictors.

(c) ...among models with at least one categorical predictor, and the same transformation of all variables.

(d) ***...among models with the same number of predictors, and the same transformations of all variables.

(e) ...when all potential predictors are independent, and we have the same transformations of all variables.

Part II: Long Answer

6. We are interested in looking again at predicting y = final year exam grades for students in Portugese schools. This time, all zero exam grades have been dropped (and the new dataset names "nozero"). The full model below considers predictors `goout` (frequency of going out with friends), `internet` (an indicator variable for having internet access at home), `studytime` (weekly study time), `Fedu` (father's education level), `health` (current health status), and `absences` (number of school absences). Output from summary statistics and variable reduction procedures are located in the appendix. The full model is:

$$y_i = \beta_0 + \beta_1 \texttt{goout}_i + \beta_2 \texttt{internet}_i + \beta_3 \texttt{studytime}_i + \beta_4 \texttt{Fedu}_i + \beta_5 \texttt{health}_i + \beta_6 \texttt{absences}_i + e_i$$

(a) Which model is chosen by each of the four model selection procedures? Be sure to specify which variables are included in the chosen models. (8 points)

AIC, AICC, and R^2_{adj} all choose the full model; AIC and AICC and minimized, while R^2_{adj} is maximized for that model. However, BIC chooses the 4-variable model with predictors `goout`, `internet`, `Fedu`, and `absences`.

(b) For each of the models chosen above (that is, with the same number of predictors), does LASSO choose the same model? If not, which model is chosen? (6 points)

Yes, LASSO chooses the same models that the other methods choose; for step 5 of the LASSO algorithm, we see the non-zero slopes for `goout`, `internet`, `Fedu`, and `absences` as above. Of course the full model is still the full model.

7. A randomized trial was conducted to investigate the relationship between a continuous response y and four treatments A, B, C, and D. The sample size was $n = 40$, with 10 observations in each of the four treatment groups. Let y be the 40×1 vector of response values, ordered so that the first 10 entries are for treatment A, the next 10 for B, then C, and finally D. The regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ was fit, where $\mathbf{X}$ is the 40×4 design matrix given by

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and where each entry is a column vector of length 10. The estimated regression coefficients were $\hat{\boldsymbol{\beta}} = [21.5, 1.3, 6.4, 8.1]'$, and residual standard deviation $\hat{\sigma} = 4.087$. Also,

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.1 & -0.1 & -0.1 & -0.1 \\ -0.1 & 0.2 & 0.1 & 0.1 \\ -0.1 & 0.1 & 0.2 & 0.1 \\ -0.1 & 0.1 & 0.1 & 0.2 \end{bmatrix}$$

(a) Express each of the four regression parameters in terms of treatment group means. For example, is $\mu_{\text{A}} = \beta_0$? (8 points)

$$\begin{aligned}
\beta_0 &= \mu_D \\
\beta_1 &= \mu_A - \mu_D \\
\beta_2 &= \mu_B - \mu_D \\
\beta_3 &= \mu_C - \mu_D
\end{aligned}$$

(b) Groups C and D used the new trial medications. Develop a hypothesis test for whether the average response for groups A and B equals the average response for groups C and D. Show all your work, especially your contrast matrix (**A**) and your hypotheses (both of them). Calculate your F-statistic. (The p-value is large.) (10 points)

Our hypotheses are $H_0 : \mu_A + \mu_B = \mu_C + \mu_D$ and $H_a : \mu_A + \mu_B \neq \mu_C + \mu_D$ (Notice that we can multiply through by 2 to remove the 1/2). The null hypothesis can be equivalently written as $\beta_1 + \beta_2 + 2\beta_0 = \beta_3 + 2\beta_0$. Writing $\mathbf{A} = [0, 1, 1, -1]'$, we can therefore write $H_0 : \mathbf{A}\boldsymbol{\beta} = 0$ vs. $H_a : \mathbf{A}\boldsymbol{\beta} \neq 0$. Now, we're working toward an F-statistic for contrasts. The denominator is the MSE for our model: $\sigma^2 = 4.087^2$. We have $\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}' = 0.4$ and $\mathbf{A}\hat{\boldsymbol{\beta}} = 1.3 + 6.4 - 8.1 = -0.4$, and the rank of $\mathbf{A}$ is $r = 1$. We therefore have

$$\begin{aligned}
F &= \frac{\left(\mathbf{A}\hat{\boldsymbol{\beta}} - h\right)\left[\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}'\right]^{-1}\left(\mathbf{A}\hat{\boldsymbol{\beta}} - h\right)/r}{RSS/(n-p-1)} \\
&= \frac{-0.4(0.4)^{-1}(-0.4)}{4.087^2} \\
&= 0.0239
\end{aligned}$$

which is not actually significant, so we don't have evidence that the average of the first two means differs from the average of the last two.

Appendix

Students Model Output

```
> cor(X)
                  goout    internet   studytime   FathersEdu       health     absences
goout       1.000000000  0.09541307 -0.04789071  0.042473845 -0.009575885  0.056589845
internet    0.095413070  1.00000000  0.07712187  0.131019467 -0.044221863  0.101567415
studytime  -0.047890707  0.07712187  1.00000000 -0.028630932 -0.072786338 -0.074541261
FathersEdu  0.042473845  0.13101947 -0.02863093  1.000000000  0.009126800  0.008947755
health     -0.009575885 -0.04422186 -0.07278634  0.009126800  1.000000000 -0.029116327
absences    0.056589845  0.10156742 -0.07454126  0.008947755 -0.029116327  1.000000000
```

```
> X <- cbind(nozero$goout, nozero$internet , nozero$studytime , nozero$Fedu,
>         nozero$health, nozero$absences)
> b<-regsubsets(X, exam)
> summary(b)

         goout internet studytime FathersEdu health absences
1  ( 1 ) " "   " "      " "       " "        " "    "*"
2  ( 1 ) "*"   " "      " "       " "        " "    "*"
3  ( 1 ) "*"   " "      " "       "*"        " "    "*"
4  ( 1 ) "*"   "*"      " "       "*"        " "    "*"
5  ( 1 ) "*"   "*"      "*"       "*"        " "    "*"
6  ( 1 ) "*"   "*"      "*"       "*"        "*"    "*"
```

Number of Predictors	R^2_{adj}	AIC	AICC	BIC
1	0.043	823.068	823.238	830.823
2	0.068	814.663	814.834	826.296
3	0.093	805.649	805.820	821.160
4	0.108	800.962	801.133	820.351
5	0.115	799.156	799.327	822.423
6	0.118	798.612	798.783	825.756

```
>X <- cbind(nozero$goout, nozero$internet , nozero$studytime ,
>             nozero$Fedu, nozero$health, nozero$absences)
> coef(mlasso)

           goout   internet studytime FathersEdu     health    absences
[1,]  0.00000000 0.00000000 0.0000000  0.0000000  0.0000000  0.00000000
[2,]  0.00000000 0.00000000 0.0000000  0.0000000  0.0000000 -0.01493730
[3,] -0.04296897 0.00000000 0.0000000  0.0000000  0.0000000 -0.02066175
[4,] -0.15860599 0.00000000 0.0000000  0.1233751  0.0000000 -0.03548667
[5,] -0.16576779 0.02130792 0.0000000  0.1294458  0.0000000 -0.03641849
[6,] -0.27366090 0.33118988 0.1121307  0.2278698  0.0000000 -0.05009852
[7,] -0.53193321 1.04222095 0.3563380  0.4646190 -0.1819079 -0.08354601
```