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

1. We have data on per-capita gross domestic product (GDP) for n countries. Let y_i be the per-capita GDP for country i, based on a population of size m_i. We are concerned about unequal variances in our regression model. What weight w_i would you use in a weighted regression model?

   (a) **** $w_i = m_i$
   (b) $w_i = \sqrt{m_i}$
   (c) $w_i = 1/m_i$
   (d) $w_i = 1/\sqrt{m_i}$

2. There is output from a multiple regression model for Amazon books in the Appendix, where the response is the price on Amazon, and there are predictor variables for each of list price, number of pages, publication year, height, and width. What null hypothesis is the F test at the bottom of the output testing?

   (a) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$
   (b) **** $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
   (c) $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$
   (d) $H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

3. Suppose x_1 and x_2 have correlation = 1. Also suppose that the correlation between x_1 and y is between 0 and 1. What will the added variable plot for x_2 look like, and why?

   (a) A horizontal line at 0, since there is no error when regressing y onto x_2.
   (b) A vertical line at 0, since there is no error when regressing y onto x_2.
   (c) A horizontal line at 0, since there is no error when regressing x_1 onto x_2.
   (d) **** A vertical line at 0, since there is no error when regressing x_1 onto x_2.
   (e) A point at 0, since there is no error in the model.

4. A researcher is interested in predicting crime rate using the following predictors: percent of each county aged 65 and up (`Pct65Up`), the percent of males who are divorced (`MalePctDiv`), and the percent of the county with public assistance (`pctWPubAsst`). Output from the Box-Cox transformation to multivariate normal are below; what transformations are suggested?

```
             Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Pct65Up         0.5573   0.1074           0.3467           0.7679
MalePctDiv      0.6043   0.1334           0.3427           0.8658
pctWPubAsst     0.0064   0.0656          -0.1222           0.1350

Likelihood ratio tests about transformation parameters
                                      LRT df         pval
LR test, lambda = (0 0 0)       49.1979101  3 1.183847e-10
LR test, lambda = (1 1 1)      242.7339877  3 0.000000e+00
LR test, lambda = (0.5 0.5 0)    0.8998975  3 8.254525e-01
```

(a) No transformations of all three variables.

(b) Log transformations of all three variables.

(c) Square root transformations of all three variables.

(d) **** Square root transformations of the first two variables and a log transformation of the third.

(e) Square root transformations of the first two variables and no transformation of the third.

5. Suppose a model with several predictors has an observed VIF of 7 for variable x_1. What does that mean in context?

(a) The correlation between x_1 and the response variable y is 7 times higher than the correlation between any of the other variables and y.

(b) The squared correlation between x_1 and y is 7 times higher than the R^2 from regressing y onto all of the other predictors.

(c) **** The variance of $\hat{\beta}_1$ is 7 times higher than it would have been if all the variables were independent.

(d) The variance of $\hat{\beta}_1$ is 7 times higher than it would have been if all the other variables had been removed.

Part II: Long Answer

6. We are interested in modeling Amazon's price (`Amazon.Price`) on a collection of books in terms of list price (`List.Price`), number of pages (`NumPages`), publication year (`Pub.year`), height (`Height`), and width (`Width`). The full model is:

$$\text{Amazon.Price} = \beta_0 + \beta_1 \text{List.Price} + \beta_2 \text{NumPages} + \beta_3 \text{Pub.year} + \beta_4 \text{Height} + \beta_5 \text{Width} + e$$

Some `R` output is shown in the appendix.

(a) Is multicollinearity a problem for this model? Why or why not? Give at least two arguments in favor of your answer. (8 points)

MULTICOLLINEARITY DOES NOT APPEAR TO BE A PROBLEM. THE PAIRS PLOT DOES NOT SHOW ANY STRONG LINEAR TRENDS BETWEEN THE PREDICTOR VARIABLES. SIMILARLY, THERE ARE NOT ANY HIGH CORRELATIONS BETWEEN THE PREDICTOR VARIABLES. NONE OF THE VIFS ARE ABOVE OUR CUTOFF OF 5. THE OVERALL F TEST IS HIGHLY SIGNIFICANT. IF NONE OF THE P-VALUES FOR THE INDIVIDUAL BETA PARAMETERS WERE SMALL, IT COULD BE EVIDENCE OF MULTICOLLINEARITY, BUT THIS IS NOT THE CASE. WE CAN ALSO LOOK AT THE ADDED VARIABLE PLOT. FLAT CURVES COULD BE INDICATIVE OF VARIABLES NOT ADDING MUCH ONCE THE OTHER VARIABLES ARE IN THE MODEL. IN THIS CASE, SINCE THE PAIRS PLOT AND CORRELATION MATRIX DO NOT SUGGEST RELATIONSHIPS WITH THE RESPONSE FOR THE VARIABLES WITH FLAT ADDED VARIABLE PLOTS, IT IS SAFE TO ASSUME THAT THE FLAT ADDED VARIABLE PLOTS ARE NOT PROBLEMATIC.

(b) Interpret the marginal model plots in Figure 3. Does our model appear to be valid? Why or why not? Also, what do the two curves in each plot represent (how were they computed)? (8 points)

The marginal model plots look good, based on the two curves overlapping to a great extent in each plot. There is a slight discrepancy between the curves at the extreme right in the plot for `List.Price`, but it is pretty minor. Marginal model plots compare nonparametric curves of y versus the predictor and $\hat{y}$ versus the predictor. Because of the expected value results on slides 21-23 of the Chapter 6 notes, we expect the two curves to be equal if the model is correctly specified.

(c) Which model is chosen by each of R^2_{adj}, AIC, AIC_C, and BIC? Be sure to specify which variables are included in each model. You can assume that each method selects the same variables for a model of the same size (e.g., for a model with 4 variables, each method can be assumed to have chosen the same 4 variables to include in the model). (8 points)

R^2, AIC, and AIC_C all choose the model with 3 variables (`List.Price`, `Pub.year`, and `Height`). BIC chooses the model with 2 variables (`List.Price` and `Pub.year`). It is not surprising that BIC chooses a simpler model, as that is its tendency.

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

For the model with 2 variables, LASSO chooses `List.Price` and `Pub.year`, which matches that chosen by the other methods. For the model with 3 variables, LASSO chooses `List.Price`, `Pub.year`, and `Height`, which again matches the other methods.

Appendix

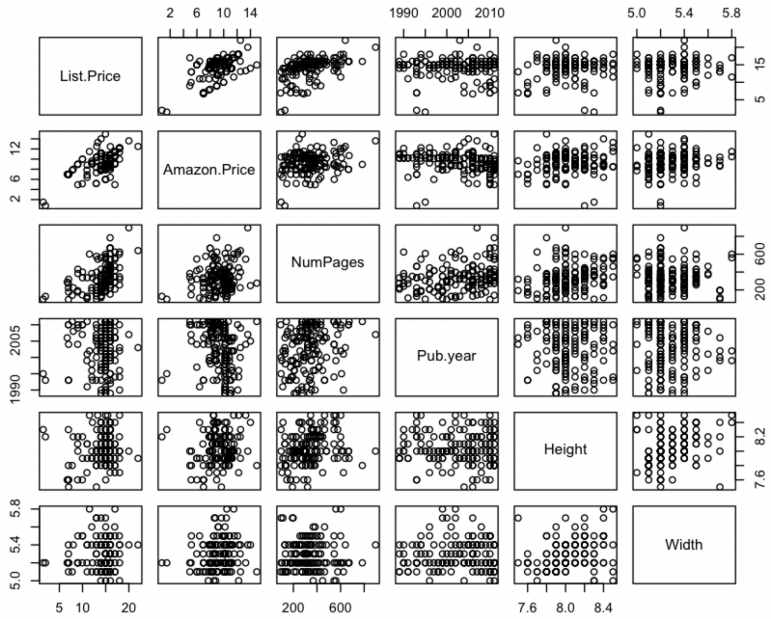

Figure 1: Pairs plot for Amazon book dataset.

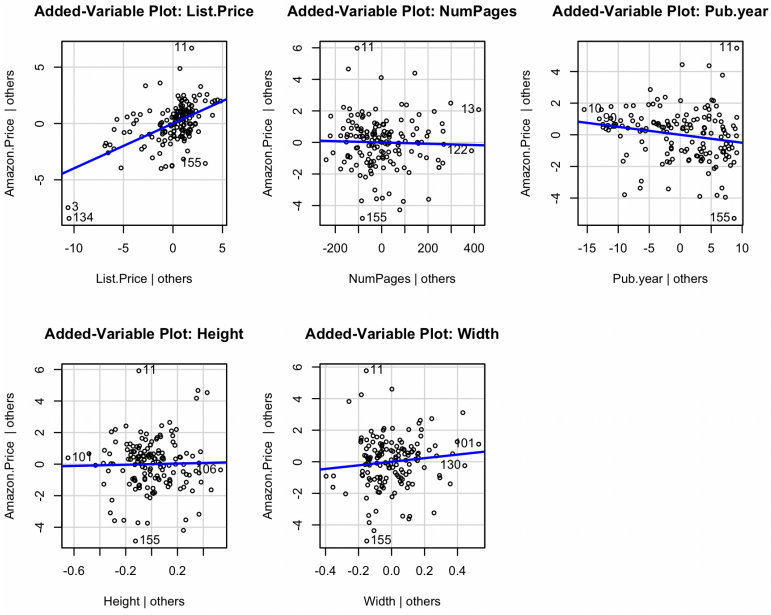

Figure 2: Added variable plots for Amazon book dataset.

```
> fit <- lm(Amazon.Price ~ List.Price + NumPages + Pub.year + Height + Width,
  data = dta)
> summary(fit)

Call:
lm(formula = Amazon.Price ~ List.Price + NumPages + Pub.year +
    Height + Width)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3591 -0.8295  0.0933  0.8371  6.1383

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.210e+02  4.037e+01   2.998   0.0032 **
List.Price   2.709e-01  5.648e-02   4.797 3.96e-06 ***
NumPages    -8.133e-04  1.059e-03  -0.768   0.4437
Pub.year    -6.284e-02  1.986e-02  -3.164   0.0019 **
Height       9.714e-01  6.495e-01   1.496   0.1369
Width        5.383e-01  8.195e-01   0.657   0.5123
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.54 on 145 degrees of freedom
Multiple R-squared:  0.251,Adjusted R-squared:  0.2252
F-statistic: 9.717 on 5 and 145 DF,  p-value: 5.063e-08

> pairs(dta)
> round(cor(dta), 3)
             List.Price Amazon.Price NumPages Pub.year Height  Width
List.Price        1.000        0.408    0.389   -0.081  0.242 -0.008
Amazon.Price      0.408        1.000    0.070   -0.284  0.221  0.086
NumPages          0.389        0.070    1.000    0.206  0.288 -0.032
Pub.year         -0.081       -0.284    0.206    1.000 -0.031  0.021
Height            0.242        0.221    0.288   -0.031  1.000  0.337
Width            -0.008        0.086   -0.032    0.021  0.337  1.000
> vif(fit)
List.Price   NumPages   Pub.year     Height      Width
  1.297058   1.445527   1.086292   1.204257   1.130236

> summary(regsubsets(Amazon.Price ~ List.Price + NumPages + Pub.year + Height + Width,
  data = dta))

         List.Price NumPages Pub.year Height Width
1  ( 1 ) "*"        " "      " "      " "    " "
2  ( 1 ) "*"        " "      "*"      " "    " "
3  ( 1 ) "*"        " "      "*"      "*"    " "
4  ( 1 ) "*"        "*"      "*"      "*"    " "
5  ( 1 ) "*"        "*"      "*"      "*"    "*"
```

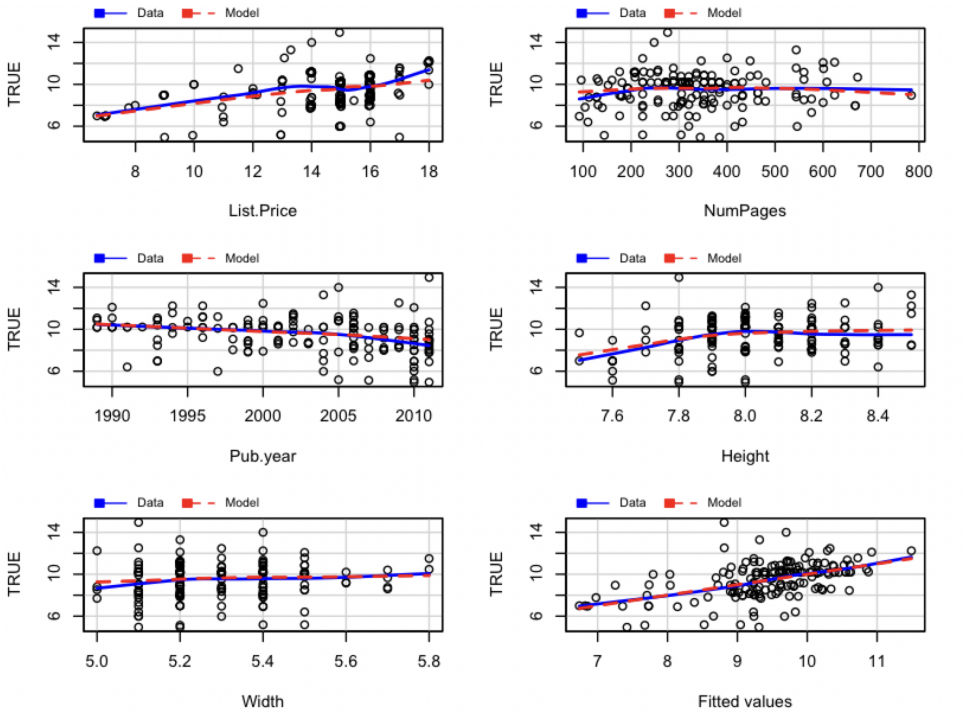

Figure 3: Marginal model plots for Amazon data

p	R^2_{adj}	AIC	AIC_C	BIC
1	0.1605	142.3858	142.5491	155.4377
2	0.2193	132.4160	132.6876	148.4827
3	0.2295	131.3930	131.8068	150.4794
4	0.2282	132.6279	133.2112	154.7315
5	0.2252	134.1792	134.9624	159.3001

Table 1: Model summaries for each value of p

p	List.Price	NumPages	Pub.year	Height	Width
1	0.02553337	0.00000000	0.00000000	0.00000000	0.00000000
2	0.16812540	0.00000000	-0.02759775	0.00000000	0.00000000
3	0.21533859	0.00000000	-0.04973213	0.57182469	0.00000000
4	0.23232328	0.00000000	-0.05711505	0.69056265	0.25486005
5	0.26954490	-0.00078330	-0.06264946	0.96127950	0.52853580

Table 2: LASSO coefficients for each value of p