**1.** **Lung pressure**. **Increased arterial blood pressure in the lungs frequently leads to the development of heart failure in patients with chronic obstructive pulmonary disease (COPD). The standard method for determining arterial lung pressure is invasive, technically difficult, and involves some risk to the patient. Radionuclide imaging is a noninvasive, less risky method for estimating arterial pressure in the lungs. To investigate the predictive ability of this method, a cardiologist collected data on 19 mild-to-moderate COPD patients. The data that follow on the next page include the invasive measure of systolic pulmonary arterial pressure and three potential noninvasive predictor variables. Two were obtained by using radionuclide imaging—emptying rate of blood into the pumping chamber or the heart and ejection rate of blood pumped out of the heart into the lungs —and the third predictor variable measures a blood gas .**

1. Prepare separate dot plots for each of the three predictor variables. Are there any noteworthy features in these plots? Comment.

A graph of a person with histogram

Description automatically generatedA graph of a graph showing the amount of blood gas

Description automatically generatedA graph of a graph

Description automatically generated with medium confidence

One of the predictor variables, blood ejection rate, seems to skew to the right. But other than that, all predictor variables more or less follow a normal distribution.

1. Obtain the scatter plot matrix. Also obtain the correlation matrix of the variables. What do the scatter plots suggest about the nature or the functional relationship between and each of the predictor variables? Are any serious multicollinearity problems evident? Explain.

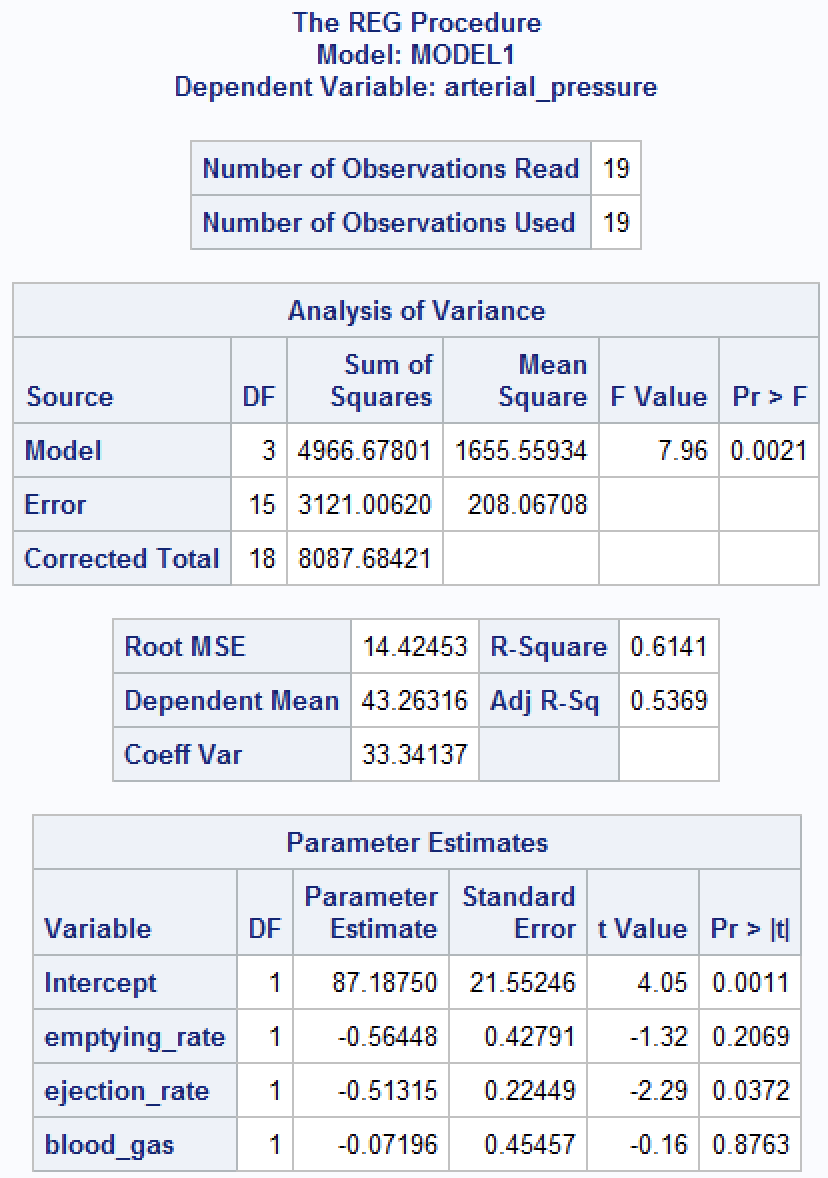
A graph of blood gas

Description automatically generatedA screenshot of a graph

Description automatically generated

There seems to be a slight positive relationship between arterial pressure vs blood gas, and a strong negative relationship for both arterial pressure vs blood emptying rate and blood ejection rate. Also, there might be issues of multicollinearity between blood emptying rate and blood ejection rate, as evidenced by the scatter plot matrix and the moderately-high correlation values between them (-0.74757).

1. Fit the multiple regression function containing the three predictor variables as first-order terms. Does it appear that all predictor variables should be retained?



The fitted regression function is . It does not appear that all predictor variables should be retained—as evidenced by the p-values, emptying\_rate and blood\_gas may be dropped.

**2. Refer to Lung pressure** **Problem 9.13.**

1. Using first-order and second-order terms for each of the three predictor variables (centered around the mean) in the pool of potential variables (including cross products of the first order terms), fitted the three best hierarchical subset regression models according to the criterion.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

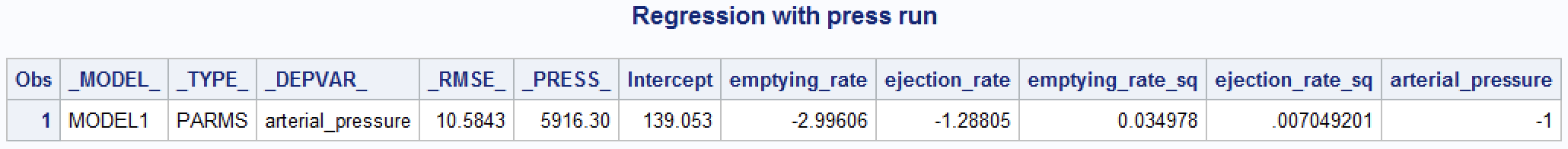
1. Is there much difference in for the best three subset models?

There is not much difference between these three models as evidenced by their similar and adjusted values.

**3. Refer to Lung pressure Problems 9.13 and 9.14. The validity of the regression model identified as best in Problem 9.14a is to be assessed internally.**

1. Calculate the statistic and compare it to . What does this comparison suggest about the validity of as an indicator of the predictive ability of the fitted model?

We will use the model with variables emptying\_rate, ejection\_rate, emptying\_rate\_sq, and ejection\_rate\_sq.



A screenshot of a computer

Description automatically generated

The of this model is 5916.30, which is much greater than its of 1568.39, so we probably shouldn’t use as a predictor.

1. Case 8 alone accounts for approximately one-half of the entire PRESS statistic. Would you recommend modification of the model because of the strong impact of this case? What are some corrective action options that would lessen the effect of case 8? Discuss.

It is reasonable to create a new dataset omitting case 8 as an outlier due to its outsized impact on the PRESS statistic. Then we would determine another best model using all subsets selection on this new dataset.

**4. The true quadratic regression function is . The fitted linear regression function is for which and What are the bias and sampling error components of the mean squared error for and for ?**

A math equations on a graph paper

Description automatically generated

**5. Refer to Lung pressure Problems 9.13 and 9.14. The subset regression model containing first-order terms for and and the cross-product term is to be evaluated in detail.**

1. Obtain the residuals and plot them separately against and each of the three predictor variables. On the basis of these plots. should any further modification of the regression model be attempted?

A screenshot of a graph

Description automatically generatedA screen shot of a graph

Description automatically generated

No further modifications seem necessary, as the residual plots do not display any unusual behavior.

1. Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be reasonable here?

A graph of a line

Description automatically generated

The residuals seem to deviate from the line at the tails, but this is to be expected since we have a small dataset. Other than that, they seem to roughly follow the line, which supports our normality assumption.

1. Obtain the variance inflation factors. Are there any indications that serious multicollinearity problems are present? Explain.

A screenshot of a computer

Description automatically generated

We know is just the inverse of tolerance, so , , and . Since we have two values greater than 10, there is evidence that serious multicollinearity problems are present.

1. Obtain the studentized deleted residuals and identify any outlying observations. Use the Bonferroni outlier test procedure with. State the decision rule and conclusion.

A screenshot of a spreadsheet

Description automatically generated



The Bonferroni outlier test statistic is . If the absolute value of any studentized statistic greater than this value, then we conclude the observation associated with the statistic is an outlier. Since no studentized statistic exceeds , we conclude there are no outliers in this dataset.

1. Obtain the diagonal elements of the hat matrix. Using the rule of thumb in the text, identify any outlying observations. Are your findings consistent with those in Problem 9.13a? Should they be? Discuss.

A screenshot of a spreadsheet

Description automatically generated



The average value is given by . If a value exceeds this average, then the observation associated with it is considered influence. Then, by the scatterplot matrix, observations 3, 8, and 15 are influential. This is indeed consistent with our conclusions in Problem 9.13a, where we observed slight skew in the histogram.

1. Cases 3, 8, and 15 are moderately far outlying with respect to their values, and case 7 is relatively far outlying with respect to its value. Obtain DFFITS, DFBETAS, and Cook's distance values for these cases to assess their influence. What do you conclude?

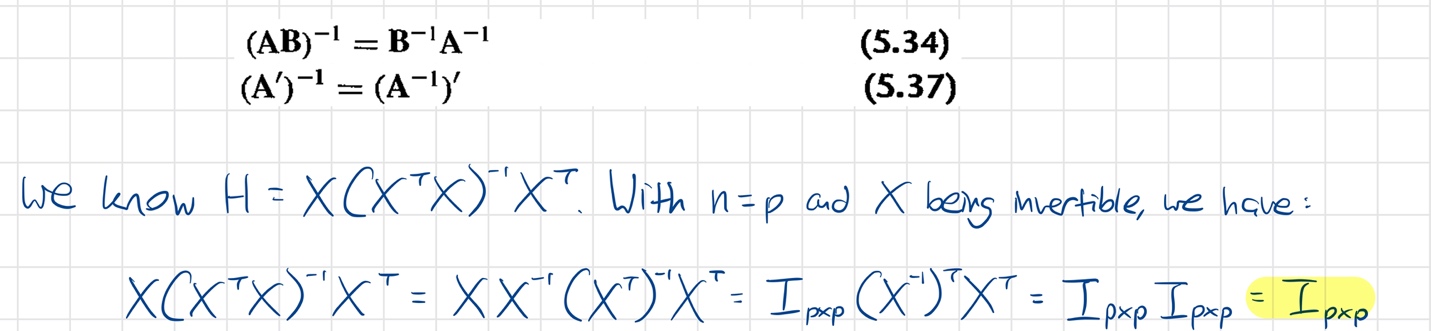
A screenshot of a spreadsheet

Description automatically generated



We arrive at the same conclusions.

**6. If and the matrix is invertible, use (5.34) and (5.37) to show that the hat matrix is given by the identity matrix. In this case, what are and .**



**7. Prove (9.11), using (10.27) and Exercise 5.31.**

