**HOMEWORK 8**

*Multiple Regression: Partial F Tests, Model Selection, Outliers, Intervals*

Answer the following questions from your textbook, *STAT2: Building Models for a World of Data*. This assignment focuses on content from Chapters 2- 4:

1. Chapter 3.6 and 4.4: Testing multiple variables (nested F-test)
2. Chapter 4.2: Model selection
3. Chapter 4.9: Outliers
4. Chapter 2.4: Confidence intervals for means and Prediction intervals

For this assignment you will use a dataset described in exercise 4.13 – North Carolina births. Use the exercise described in the book on page 191 along with the dataset posted on Canvas to answer the following questions. Note that the sample size in the dataset provided to you is smaller than the book because rows with missing data have been excluded. You do not need to answer the questions listed in the book.

For all questions, the response variable will be the birthweight in ounces (BirthWeightOz).

1. For this question, you will consider the following explanatory variables in the .jmp file:

* Completed weeks of gestation (Weeks)
* Mother’s race (MomRace coded as black, Hispanic, white, or other; You can choose the baseline/reference group.)

Start by creating indicator variables for mother’s race and then use Fit Model to fit a model to predict birthweight using weeks of gestation and mother’s race.

* 1. Conduct a Partial F test (custom test in JMP) to test mother’s race. Include only the output from the Custom Test in JMP.

Ho: β3= β4= β5=0 where β3= slope for other race, β4= slope for Hispanic race, and β5= slope for black race

Ha: At least one βi *≠* 0

Test statistic = F-statistic = 8.401

*p*-value = 0.0000155

Conclusion: There is overwhelming evidence to suggest that Mother’s Race is a predictor of the birthweight in ounces.

Table

Description automatically generated

1. For this question, you will consider the following explanatory variables in the .jmp file:

* Completed weeks of gestation (Weeks)
* Mother’s age in years (MomAge)
* Weight gained during pregnancy in pounds (Gained)
* Sex of the baby (Sex; female = 0 = baseline/reference group)
* Indicator for premature birth (Premie; no = 0 = baseline/reference group)

Fit the following 4 models and fill in the table below. You DO NOT need to include your JMP output for fitting all of these models. Round all numbers to 4 decimal places.

Model 1: Weeks

Model 2: Weeks and MomAge

Model 3: Weeks, MomAge, and Gained

Model 4: Full model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Model 1 | Model 2 | Model 3 | Model 4 |
| Number of Explanatory Variables | **1** | **2** | **3** | **5** |
| Adjusted R2 | **0.3419** | **0.3591** | **0.3829** | **0.3999** |
| RMSE | **17.9488** | **17.7126** | **17.3815** | **17.1399** |
| Mallow’s Cp | **137.9219** | **98.5210** | **43.8840** | **6** |
| Overall  F-test *p*-value | **< 0.0001** | **< 0.0001** | **0.3268** | **0.8147** |
| Largest *p*-value for all individual coefficient *t*-tests: vs | **< 0.0001** | **< 0.0001** | **< 0.0001** | **0.0002** |

* 1. Choose one of the models from the table. Make sure to comment on all characteristics listed in the table.

Based on the Adjusted *R*2 criterion and number of variables, Model 4 is the best as it has the highest adjusted R2. Based on Mallow’s Cp criterion and number of variables, Model 4 is the best as it has the lowest Mallow’s Cp. Based on the RMSE criterion and the number of variables, Model 4 is the best as it has the lowest RMSE. Based on the F-test criterion and number of variables, Model 1 is the best as it has the lowest p-value and the smallest number of variables. Based on the p-values for the hypothesis test *βi* = 0 vs *βi ≠* 0 and the number of variables, Model 1 is the best. It has 1 significant p-value and has the smallest number of variables.

The above tools show that Model 4 or Model 1 is the best. I am choosing Model 1 as the best because it has the smaller number of variables which makes it a simpler model.

1. For this question, you will work with the full model you fit in the previous question. Create a leverage and influence plot. Do you have any concerns with outliers? Make sure to discuss standardized residuals, leverage, and Cook’s D. Specify the cutoffs you used to determine if the outlier was extreme. Include a copy of your leverage plot with your assignment. Your plot should include horizontal lines to help identify extreme standardized residuals and vertical lines to help identify high leverage values.

k=5, n = 1409

High Leverage: h > 2((1+5/1409) = h > 2(6/1409) = h > 0.00852

Very High Leverage: h > 3((1+5)/1409) = h > 3(6/1409) = h > 0.01278

Chart, scatter chart

Description automatically generated

There are 2 points with very high standardized residuals. The two points are above 3 and are shown as the red points. There are approximately 22 points with very high leverage values above 0.01278. There are no points with high influence with Cook’s D values above 0.5 or very high influence points with Cook’s D values above 1.

1. For this question, continue to work with the full model you fit in question 2. Consider the first observation in the dataset where Sex=male, MomAge=32, Weeks=40, Gained=38, and Premie=no. Compute and interpret a confidence interval for the mean birthweight and a prediction interval for the first observation in the data. You do not need to include any output from JMP.

Confidence Interval: [127.30743999, 130.53088101]

We are 95% confident that the mean birthweight in ounces for a male sex baby, Mom’s age of 32 years, 40 weeks of gestation, 38 pounds of weight gained during pregnancy in pounds, and born not premature, is between 127.31 and 130.53 ounces.

Prediction Interval: (95.257877745, 162.58044326)

We are 95% confidence that the birthweight in ounces for a male sex baby, Mom’s age of 32 years, 40 weeks of gestation, 38 pounds of weight gained during pregnancy in pounds, and born not premature is between 95.26 and 162.58 ounces.