Print

Practice: Regression Modeling Mini Case Study

Open the course data set **Bodyfat.jmp** in JMP. (Note that a version of this file is available in the JMP Sample Data Library, but you will use the course file for this practice.) Recall that the data set consists of **%Fat** and other physical measurements from 252 men.

In this exercise, you conduct a regression analysis for **%Fat** as a function of the other variables. You'll explore the variables, fit a full model, explore multicollinearity, conduct a residual analysis, check for influential values, and develop a reduced model for **%Fat**.

Notice that the purpose of this exercise is to provide you with an opportunity to practice applying what you learned throughout this lesson. Show the solutions after each question to check your work. In a real-life modeling situation, you might make different decisions regarding the best course of action at each step.

1. Create histograms for all of the variables. Are there any missing values? Are there any unusual observations that should be hidden or excluded from the analysis?

   There are no missing values for any of the variables. There is a severe outlier for **Height**. **Height** is reported in inches, and it is highly unlikely that a man is 29.5 inches and weighs 205 pounds. There was likely a typographical error, so this observation should be hidden and excluded from the analysis. For the remainder of this exercise, this outlier is hidden and excluded. There are other outliers, but the range of values for all the variables looks reasonable.

2. Create a scatterplot for all of the variables.

   a. Is the response, **%Fat**, highly correlated with any of the predictors (with a correlation coefficient of |+/- 0.7| or higher)?

   b. Are any of the predictors highly correlated with one another? (There are many pairs of variables, so use the scatterplot matrix to generally assess correlations.)

   c. Are there any unusual patterns or observations in the scatterplot matrix?

   a. Select the **Multivariate** platform on the **Analyze, Multivariate Methods** menu. **Abdomen** (0.8126), **BMI** (0.7249), and **Chest** (0.7029) are highly correlated with **%Fat**.

   b. There are many correlations between predictors. **Weight** and **BMI** seem to be most strongly correlated with the other predictors and with one another.

   c. Some points seem to be outliers for many of the variables.

3. Fit a full model, with **%Fat** as the response and all the other variables as predictors. Which two variables are the most significant in this model?

   Abdomen (*p*-value = 0.00000) and **Wrist** (*p*-value = 0.00130) are the most significant.

4. Use VIFs to investigate multicollinearity in this model. To requests VIFs, right-click the **Parameter Estimates** table and select **Columns** and then **VIF**.

   a. Overall, do the VIFs indicate that there is a problem with multicollinearity?

   b. For this exercise, remove **BMI** from the model. (Notice that a simplistic approach is taken for the purposes of this exercise.) What happens to the VIFs for the other variables after **BMI** is removed?

   a. Yes, three of the VIFs are very high. The VIF for **Weight** is 253, the VIF for **BMI** is 218, and the VIF for **Height** is 54.

   b. The VIF for **Weight** is still high (44.7), but the VIF for **Height** (2.93) is now acceptable. There are three other VIFs greater than 10.

5. Look at the residual plots. Are there any unusual patterns or observations?

   There are no unusual patterns or observations in the residual plots. The residuals in the Residual by Predicted plot are randomly scattered about zero, the residuals are approximately normally distributed, and the Studentized Residual plot doesn't show any outliers.

6. Save **Cook's D Influence** values to the data table and use the **Distribution** platform to graph the values. Are there any influential values that you should be concerned with?

   There are one or two observations that have extreme Cook's D values relative to the other observations, but none of the Cook's D values is greater than 1.0.

7. The initial "Full" model has all the predictors except **BMI**.

   Hint: For these questions, you might need to select the **Summary of Fit** and **Analysis of Variance** tables from the top red triangle under **Regression Reports**.

   a. Is this model significant?

   b. What are the RSquare Adjusted and RMSE for this full model?

a. Yes, the *p*-value in the ANOVA table is < 0.0001.

b. RSquare Adjusted is 0.732 and RMSE is 4.31.

8. Use the Effect Summary table to slowly remove nonsignificant terms from the model, and use a cutoff *p*-value of 0.05. Stop when the largest *p*-value that remains is <0.1.

   a. How many terms are in the reduced model?

   b. What are the RSquare Adjusted and RMSE for this reduced model?

   c. Compare the fit statistics (RSquare Adjusted and RMSE) between the full model and the reduced model. Which model performs better?

   d. Assume that you use one of these two models for predictive purposes. You take several physical measurements and use the model to predict a man's **%Fat**. From a practical perspective, which model would you rather use, the larger full model or the reduced model? Why?

   a. 7, **Abdomen, Weight, Forearm, Wrist, Age, Thigh**, and **Neck**

   b. RSquare Adjusted is 0.734 and RMSE is 4.298.

   c. The statistics for the reduced model are slightly better, but they are very close.

   d. The models perform very similarly. The smaller model requires data to be collected on fewer variables and is much easier to use. The smaller model would be better.

Hide Solution

Close