Math 5366 Homework 23

1. Import the file `math5305Lab6Data.txt`, whose columns are the variables Y, X_1, X_2, and X_3. The goal of these first three problems is to perform diagnostics to assess the assumptions of normality and constancy of variance for a model predicting Y from the X_j's, to transform Y if necessary, to assess the transformed model using diagnostics, and to compare the original and transformed models via residual sums of squares.

   We begin by fitting a model and assessing it with diagnostics.

   (a) Fit `model = lm(Y~X1+X2+X3)`, compute the fitted values $\hat{Y}$, and the residuals e. You can use the command `Yhat=predict(model)` to obtain $\hat{Y}$.

   (b) Plot Y vs. $\hat{Y}$. If the model were valid, what would you expect this plot to look like? Does the plot suggest the existence of curvature in the model?

   (c) Plot e vs. $\hat{Y}$. If the model were valid, what would you expect this plot to look like? Does the plot suggest the existence of curvature in the model?

   (d) Now that we know curvature is present, there are two courses of action we can take: transform Y or transform the X_j's, or both. Generally, if there are problems with the errors, we should transform Y, and if the errors are ok, we should transform the X_j's. Let's investigate the errors.

   (e) Plot a qq-plot to check normality of the error terms using the `qqnorm` command.

   (f) Perform the Shapiro-Wilks test to check normality of the error terms using the `shapiro.test` command.

   (g) Based on the results in parts (e) and (f), do the error terms for this model appear to be normal?

   (h) Check constancy of error variance by plotting $|e|$ vs. $\hat{Y}$.

   (i) Check constancy of error variance by performing the Brown-Forsythe test.

   (j) Based on the results in parts (h) and (i), do the error terms for this model appear to have constant variance?

   (k) Does a transformation of Y appear to be necessary?

   (l) Finally, calculate the residual sum of squares $\|e\|^2$. Note that this value is

   $$\|e\|^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

   so it is similar to a prediction sum of squares (PRESS). It measures the sum of square errors between the predictions $\hat{Y}_i$ and the actual observations Y_i. This number is very large, so to put it in perspective, calculate $\frac{\|e\|^2}{\|Y-\overline{Y}\|^2}$. Assessing the model by this criterion is equivalent to using $R^2 = 1 - \frac{\|e\|^2}{\|Y-\overline{Y}\|^2}$.

2. Let λ be the optimal value produced by the Box-Cox transformation. Transform Y by defining $\tilde{Y}_i = Y_i^{\lambda}$, for $i = 1, \ldots, 100$.

   (a) Fit a model `tmodel` by regressing $\tilde{Y}$ on X_1, X_2, and X_3, and find the corresponding fitted values $\hat{\tilde{Y}}$ and $\tilde{e}$.

(b) Plot $\tilde{Y}$ vs. $\hat{\tilde{Y}}$ and $\tilde{e}$ vs. $\hat{\tilde{Y}}$. How do these plots compare to those from problem 2? Does curvature appear to exist in the transformed model?

(c) Investigate normality of the errors for the transformed model.

(d) Investigate constancy of error variance for the transformed model.

(e) Do the errors for the transformed model appear to satisfy the assumptions of normality and constant error variance? How do your results compare to those from problem 2?

3. Now let's apply the results from the transformed model to the original variable Y.

(a) First, create a vector of fitted values for Y by defining $\hat{Y}_i = \hat{\tilde{Y}}_i^{1/\lambda}$, for $i = 1, \ldots, 100$, and create a vector of residuals by defining $e_i = Y_i - \hat{Y}_i$, for $i = 1, \ldots, 100$. These are predicted values and residuals for the original model, but they take advantage of the information from the transformed model.[1]

(b) Plot Y vs. $\hat{Y}$ and e vs. $\hat{Y}$. Did the transformation appear to correct problems with the functional form? [2]

(c) Finally, calculate $\|e\|^2$, $\frac{\|e\|^2}{\|Y-\overline{Y}\|^2}$, and $R^2 = 1 - \frac{\|e\|^2}{\|Y-\overline{Y}\|^2}$ as in question 2. Which model fits the data better/has a lower residual sum of squares?[3]

4. Import the UCI Machine Learning Repository's Auto-Mpg data set, and create the best possible linear regression model for predicting `mpg` from the other variables. Use diagnostics and remedial measures to investigate curvature and assumptions related to the design matrix and error terms.

Notes

[1]The e_i's may *not* satisfy normality and homoscedasticity, because they are transformed versions of the $\tilde{e}_i$'s, which do satisfy these assumptions. The model satisfying these assumptions is the *transformed* model. However, the transformed model only allows us to predict $\tilde{Y}$, but what we really care about is $Y = \tilde{Y}^{1/\lambda}$, which is why we are estimating Y with $\hat{Y}_i = \hat{\tilde{Y}}_i^{1/\lambda}$.

[2]Since the plot of Y vs. $\hat{Y}$ hugs the line $y = \hat{y}$, we conclude that $Y \approx \hat{Y}$, that is, the $\hat{Y}$'s are doing a good job of predicting the Y's. Similarly, the lack of trend in the plot of e vs. $\hat{Y}$ indicates that the functional form has been corrected.

[3]Note that you can't use a model summary or the equation $R^2 = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}$ to calculate R^2 on this problem, because our final model is actually nonlinear due to the Box-Cox transformation.