STA 5207 Assignment 6

Due Friday October 22

1. (50 points total) Consider the data set divusa in the faraway package: Divorce rates in the USA from 1920-1996

There are 77 observations and 7 columns:

year: the year from 1920-1996

divorce: divorce per 1000 women aged 15 or more

unemployed: unemployment rate

femlab: percent female participation in labor force aged 16+

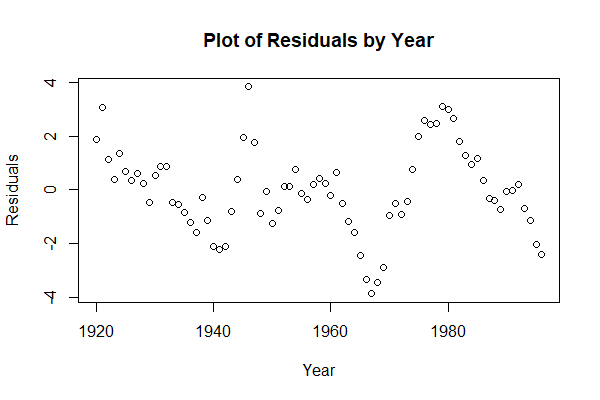
marriage: marriages per 1000 unmarried women aged 16+

birth: births per 1000 women aged 15-44

military: military personnel per 1000 population

The data set divusa.txt in SAS Studio has 8 columns, with the observation number as the first column.

Fit a regression model with divorce as the response and unemployed, femlab, marriage, birth, and military as predictors.

1. (10 points) The variable year is not being used in the model, but it shows that the measurements were taken across time. What does this make us suspect about the error terms? No output needed.
   1. We can suspect the errors are correlated.
2. (15 points) Make a plot of the residuals by year and notice the pattern. Does it appear from this that the errors are serially correlated? Submit the plot.
   1. It does appear the errors are serially correlated.
3. (10 points) Perform the Durbin-Watson test on the residuals at 5% significance level. Give the hypotheses, p-value, and conclusion.
   1. H0:errors are uncorrelated vs. H1:errors follow a first-order autoregressive process

|  |  |  |  |
| --- | --- | --- | --- |
| Lag | Autocorrelation | D-W Statistic | p-value |
| 1 | 0.8260933 | 0.2998834 | 0 |

* 1. At the 5% significance level, we conclude that the errors follow a first order autoregressive process.

1. (15 points) Model the serial correlation with an AR(1) process, (like we did in class) meaning that Σij = ρ|i-j|. Use ML method to estimate the parameters in the GLS fit. Are there differences between which predictors are significant (at 5%) in the OLS model and which are significant (at 5%) in the GLS model? If so, state the changes.

Correlation Structure: AR(1)

Formula: ~1

Parameter estimate(s):

Phi

0.9715486

Coefficients (GLS):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Value | Std.Error | t-value | p-value |
| (Intercept) | -7.059682 | 5.547193 | -1.272658 | 0.2073 |
| Unemployed | 0.107643 | 0.045915 | 2.344395 | 0.0219 |
| femlab | 0.312085 | 0.095151 | 3.279878 | 0.0016 |
| marriage | 0.164326 | 0.022897 | 7.176766 | 0.0000 |
| birth | -0.049909 | 0.022012 | -2.267345 | 0.0264 |
| military | 0.017946 | 0.014271 | 1.257544 | 0.2127 |

Call:

lm(formula = divorce ~ unemployed + femlab + marriage + birth +

military, data = divusa)

Residuals:

Min 1Q Median 3Q Max

-3.8611 -0.8916 -0.0496 0.8650 3.8300

Coefficients:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| (Intercept) | 2.48784 | 3.39378 | 0.733 | 0.4659 |
| Unemployed | -0.11125 | 0.05592 | -1.989 | 0.0505 |
| femlab | 0.38365 | 0.03059 | 12.543 | < 2e-16 |
| marriage | 0.11867 | 0.02441 | 4.861 | 6.77e-06 |
| birth | -0.12996 | 0.01560 | -8.333 | 4.03e-12 |
| military | -0.02673 | 0.01425 | -1.876 | 0.0647 |

* 1. There is a difference between which predictors are significant in the OLS and GLS model at 5% significance. Using the GLS model, unemployed becomes significant whereas it isn’t in OLS.

1. (50 points total) The data set “labor.txt” contains a response and 5 predictors. We will model “monthly labor hours” as a function of the 5 predictors. The variables are:

X1: average daily patient load

X2: monthly x-ray exposures

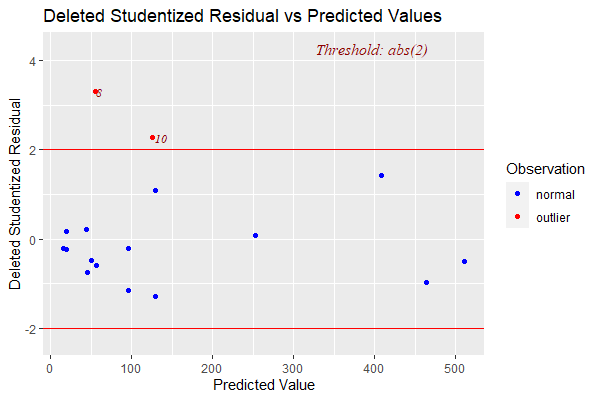
X3: monthly occupied bed days

X4: eligible patient population (/1000)

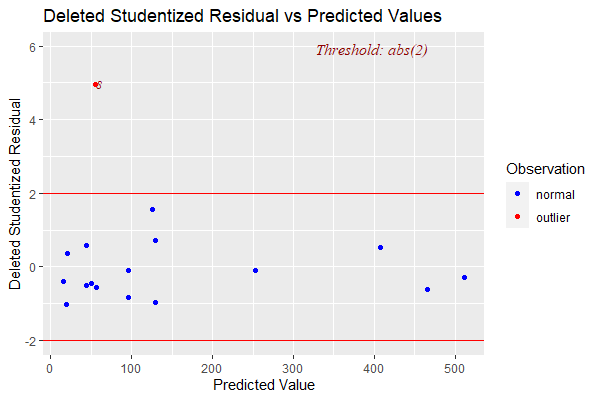
X5: average length of stay in days

Y: monthly labor hours

1. (15 points) Perform OLS regression and examine a plot of studentized residuals versus fitted values. This plot indicates nonconstant variance. What is the pattern demonstrated by this plot? What is the fitted model? Submit the plot only.



* 1. This is a nonlinear pattern with high spread
  2. y = -4.091 – 9.928e-07x1 + 0.0318x2 + 0.0489x3 + 0.937x4 – 0.000148x5

1. (10 points) It appears that the variance is increasing with the predicted values, so it may appropriate to use WLS. Based on the pattern in the studentized residual plot, what would make a good choice for the weights? No output needed.
   1. A good choice for the weights would be *wi* = 1/*xi*
2. (25 points) Perform WLS using the inverse of the predicted values as weights and examine the plot of studentized residuals versus fitted values. Does the weighted model appear to satisfy the assumption of constant variance? Submit the plot.
   1. The weighted model has much less spread which appears to satisfy the assumption of constant variance.