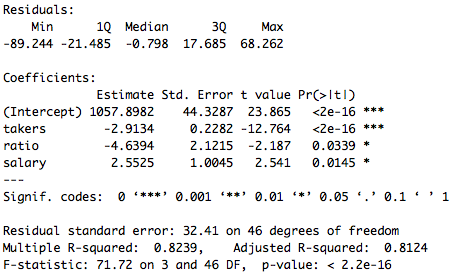
Israel Diego

2/11/16

STATS 500 Homework 3

**Part A**

Problem 1 Fitting a regression of **total** SAT score on the predictors: **takers**, **ratio**, and **salary**, we get a Multiple R-squared statistic of 0.8239. This can be considered as a good fit for the model as 82.39 % of the variation in **total** is explained by these predictors.



Problem 2 To test whether teachers’ salary has a positive effect on SAT scores, we set up a one-sided *t*-test where ,. From our regression in Problem 1, we get a *t-*statistic of 2.541 for . Using the command: pValueSalary = (1 - pt(2.541, df=46)), gives us a p-value: 0.00724543617, which is smaller than . We conclude that we reject our null hypothesis at the 1% level.

Problem 3 To test whether ratio has a positive effect on SAT scores, we set up a two-sided *t*-test where , ,. From our regression in Problem 1, we get a *t-*statistic of 2.541 for . Using the command: pValueSalary = 2\*(1 - pt(2.187, df=46)), gives us a p-value: 0.03386468, which is bigger than . We conclude that we reject our null hypothesis at the 5% level, but not at the 1%.

Problem 4 To test the , against the alternative that at least one of the estimates is not 0, we can use the F-statistic and p-value given in the output of Problem 1: 71.72 and 2.2e-16 respectively. Given this p-value is very small we conclude that we may reject the null hypothesis at virtually any of the conventional significance levels. The null hypothesis implies that none of the estimates used to explain the response variable, have any explanatory power, and thus is basically a hypothesis test on the entire regression model.

Problem 5

|  |  |  |
| --- | --- | --- |
| **C.I.** | **0.5%** | **99.5%** |
| salary | -0.146684 | 5.251624 |

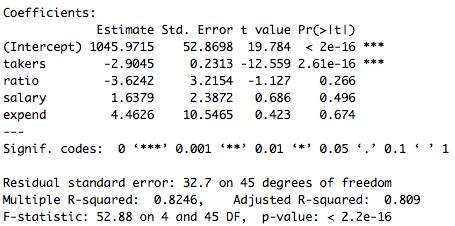
|  |  |  |
| --- | --- | --- |
| **C.I.** | **2.5%** | **97.5%** |
| salary | 0.5304797 | 4.5744605 |

From both confidence intervals we check to see if 0 lies within them. We see that 0 is not inside the 95% interval. However 0 is inside the 99% interval. We conclude that the p-value associated with must be less than .05 but greater than .01.

Macintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:figures:Conf region.pdf

Problem 6 Below is a plot of the 95% confidence region of ratio and salary. Note that the location of the origin is outside of the ellipse, which implies we reject the null hypothesis that . Now considering this origin point with respect to the vertical dotted lines, we notice that the origin lies outside of this interval, so we reject the null hypothesis that . Now if we focus on the horizontal lines, we notice that the origin lies inside the region made between the two horizontal lines. This indicates that we would fail to reject the null hypothesis that .

Problem 7 Fitting a regression of **total** SAT score on the predictors: **takers**, **ratio**, **salary**, and **expend**, we get a Multiple R-squared statistic of 0.8246. This is slightly higher than the goodness of fit from Problem 1. The estimate on takers does not change much, while the estimate on ratio is more positive by 1.0152 more. Also the estimate of ratio is not statistically significant anymore. Similarly the estimates of salary and expend are not statistically significant either, and the estimate of salary is lower than that of the estimate in problem 1 by 0.9146.



Problem 8 To test the null hypothesis, , against the alternative hypothesis that at least one of the estimates is not 0, we can use the method outlined in class:

> hO8 = lm(total ~ takers, data = sat)

> hOa8 = lm(total ~ takers + ratio + salary + expend, data = sat)

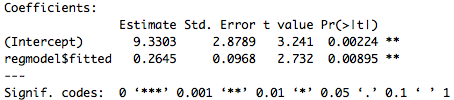
> anova(hO8,hOa8)

The result gives us an F-statistic = 3.2133 and a p-value of 0.03165. This p-value indicates that we may reject our null hypothesis at the 5% level, but not at the 1% level.

Based on this analysis, it would seem to me that the first regression from Problem 1 gave better results than the regression in Problem 7. Adding the **expend** variable in Problem 7 raised our standard errors to the point that only the **takers** variable was statistically significant. The improvement on the goodness of fit was very small as well. Thus the costs of adding the expend variable far outweigh the benefits. I would not consider the expend variable significant in our regression, and would probably look for a possibly better alternative.

Macintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:Resid against fitted.pdf**Part B**

Initially we run a regression of gamble on sex, status, income, and verbal. To check the constant variance assumption, we can first plot the residuals of this regression model against its fitted values. We notice there is a small trend present in this plot, so we can run a regression of the abs(residuals) on the fitted values. The results below demonstrate that there is a positive relationship between residuals and fitted values with a slope of 0.2645. This estimate is statistically significant at the 1% level indicated by the regression summary below. This would indicate that we have heteroscedasticity in the residuals. Thus taking the square root of the gamble variable and regressing on our original predictors will fix this.



Macintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:figures:Normal Q-Q Plot.pdfMacintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:figures:Hist of Residuals.pdf

Next is to check if the normality assumption for our regression still holds. For this we make a QQ-plot of the residuals, demonstrated on the right. The QQ- plot seems to resemble the shape of a Cauchy distribution, with fat tails. We can use the Shapiro-Wilk test of normality to check this assumption. This test gives us a p-value of 0.7272, which means we reject fail to reject the null hypothesis that residuals are normal given that the p-value is so large.

Macintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:figures:Plot of leverages.pdfAs demonstrated in class, we can construct the Half-normal plot for leverages. In this case we see from the plot on the left that the leverage points correspond to the 42nd and 35th observations. We can view these observations and see that they both correspond to male observations, however they are very distinct from each other in all other factors.

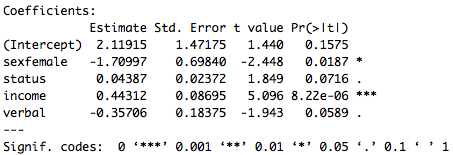
Macintosh HD:Users:rdiego:Desktop:Screen Shot 2016-02-09 at 12.12.54 AM.png

In order to find the Outliers we compute externally studentized residuals, and take the maximum *t*-statistic in absolute value, and test its p-value with (47-(4+1)-1 = 41) degrees of freedom. We also apply the Bonferroni correction which is . Our p-value comes out to 0.00414 and the critical value is 0.00106. Since the p-value is larger, we fail to reject the null hypothesis that this point is not an outlier. Therefore we conclude that none of the observations can be outliers because we failed to reject for the maximum studentized residual.

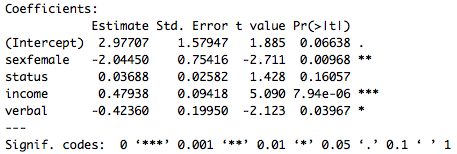
Macintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:figures:Cooks distance plot.pdf

We look for influential points in the data, which can have a large change in the fit of our model. We compute cook’s distance and plot as shown in lecture. The plot on the right shows three possible influential points corresponding to observations 5, 39, and 24.

Below is the regression summary excluding the 24th observation.



Below I also add the summary of our original regression:



We see that excluding this observation results in our sex estimate being significant at the 5% level but not the 1% anymore. The verbal estimate is not significant at the 5% level anymore. The *R2* statistic drops from 0.5646 to 0.5503.

Next we can construct Partial regression plots excluding income. We notice there is a positive relationship between the gamble and income residuals and no sign of nonlinearity. We also add a partial residual plot, shown on the bottom right, which shows there is still no significant sign of nonlinearity.

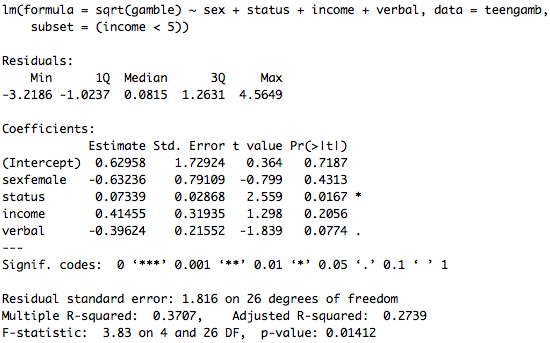
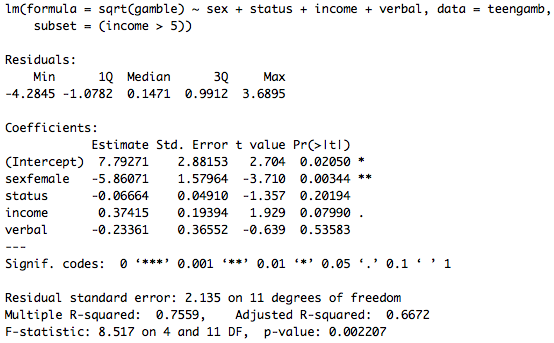
**Macintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:figures:Gamble Income residuals.pdfMacintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:figures:Gamble adj Income.pdf**

Below we show plots of residuals vs predictors. We can observe there is practically no correlation between the predictor variables and the residuals of our regression model.

**Macintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:figures:Status Residuals.pdfMacintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:figures:Income Residuals.pdfMacintosh HD:Users:rdiego:Documents:Senior Year:Winter 2016:Stats 500:homeworks:HW3:figures:Verbal Residuals.pdf**

Finally we can also run regressions on two different types of groups, individuals with weekly income below 5 pounds per week, and those with greater than 5 pounds per week. Next page we demonstrate their regression summaries.

Income > 5 pounds/week Income < 5 pounds/week



The regression on those with income greater than 5 pounds, explains about 75% of the variation in gambling expenditures, while the regression of those individuals with less than 5 pounds, explains about 37% of the variation in gambling expenditures. We can infer that individuals with the higher levels of income contribute a larger part in explaining gambling expenditures than those with lower levels.

**Part C**

Say the two predictor variables *A* and *B* are positively correlated. The goal is to explain why the confidence region of *A* and *B* will “lean to the left.” If we consider the case where we would like to estimate , it would be easier to estimate this relationship, because there will be little uncertainty about this value. In the case where we estimate , this is more difficult to estimate since there will be more uncertainty in our estimate of the difference. The uncertainty is reflected in the elongated shape of the confidence region, where the difference of the two estimates would lie. Hence the confidence region will have the shape of a downward sloping ellipse, leaning to the left.

In the case where *A* and *B* are anti-correlated, it must be the case that their confidence region will be leaning right. The intuition is the opposite of the case where *A* and *B* are positively correlated. Estimating is easier because there is greater certainty about their difference. In the case where we estimate their sum, this is more difficult to estimate since there will be more uncertainty in our estimate of the sum. The uncertainty is reflected in the elongated shape of the confidence region, where the sum of the two estimates would lie. Hence the confidence region will have the shape of an upward sloping ellipse, leaning to the right.