**Statistics 512: Solution to Homework#6**

**For the following 3 problems use the computer science data that we have been discussing in class. You can get a copy of the data set `csdata.dat` from the class website. The variables are: `id`, a numerical identifier for each student; `GPA`, the grade point average after three semesters; `HSM`; `HSS`; `HSE`; `SATM`; `SATV`, which were all explained in class; and `GENDER`, coded as 1 for men and 2 for women.**

1. In a `data` step, create a new variable `GENDERW` that has values 1 for women and 0 for men (use arithmetic on the original variable `GENDER`). Run a regression to predict `GPA` using the explanatory variables `HSM`, `HSS`, `HSE`, `SATM`, `SATV`, and `GENDERW`. (Do not include any interaction terms.)

   **Solution:** The required SAS output is given below:

   Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 0.31101 | 0.40447 | 0.77 | 0.4428 |
| genderw | 1 | 0.03237 | 0.11148 | 0.29 | 0.7718 |
| hsm | 1 | 0.14423 | 0.03979 | 3.62 | 0.0004 |
| hss | 1 | 0.03827 | 0.03874 | 0.99 | 0.3244 |
| hse | 1 | 0.05103 | 0.04228 | 1.21 | 0.2287 |
| satm | 1 | 0.00100 | 0.00071725 | 1.40 | 0.1633 |
| satv | 1 | -0.00041086 | 0.00059323 | -0.69 | 0.4893 |

   (a) Give the equation of the fitted regression line using all six explanatory variables.

   **Solution:** The estimated regression equation is

$$\begin{aligned} gpa &= 0.31101 + 0.003237 genderw + 0.11423 hsm + 0.03827 hss + \\ &\quad 0.05103 hse + 0.001 satm - 0.00041086 satv \end{aligned}$$

   (b) Give the fitted regression line for women (use part a).

   **Solution:** To get the regression line for women, set $genderw = 1$ in the line from the previous problem:

$$\begin{aligned} gpa &= 0.34388 + 0.11423 hsm + 0.03827 hss + 0.05103 hse + \\ &\quad 0.001 satm - 0.00041086 satv. \end{aligned}$$

   (c) Give the fitted regression line for men (use part a).

   **Solution:** To get the regression line for men, set $genderw = 0$ in the line from the previous problem:

$$\begin{aligned} gpa &= 0.31101 + 0.11423 hsm + 0.03827 hss + 0.05103 hse + \\ &\quad 0.001 satm - 0.00041086 satv. \end{aligned}$$

2. Use the $C_p$ criterion to select the best subset of variables for this problem (i.e. use the options " `/ selection = cp b;`") . Use only the original six explanatory variables, not `HS` or `SAT`, and use either `GENDER` or `GENDERW`, not both. Summarize the results and explain your choice of the best model.

**Solution:** The following five models satisfy the two criteria: $C_p \leq p$, and smallest $C_p$ for a given $p$.

C(p) Selection Method

| Number in Model | C(p) | R-Square | Parameter Estimates: Intercept | hsm | hss | hse | satm | satv | gender |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 1.8079 | 0.2016 | 0.62423 | 0.18265 | . | 0.06067 | . | . | . |
| 3 | 2.3303 | 0.2069 | 0.30471 | 0.16272 | . | 0.06572 | 0.00074675 | . | . |
| 4 | 3.5571 | 0.2097 | 0.27775 | 0.14951 | 0.03311 | 0.05060 | 0.00073229 | . | . |
| 5 | 5.0843 | 0.2115 | 0.32672 | 0.14596 | 0.03591 | 0.05529 | 0.00094359 | -0.00040785 | . |
| 6 | 7.0000 | 0.2118 | 0.27864 | 0.14423 | 0.03827 | 0.05103 | 0.00100 | -0.00041086 | 0.03237 |

(Interestingly, the model with only hsm is excluded because it has $C_p = 2.3292 > 2 = p$.) Any of these might be considered reasonable according to the $C_p$ criterion, except perhaps the model with 6 parameters since it has $C_p = p$ and the others have $C_p < p$.
I choose the model with only 2 parameters, hsm and hse, for the following reasons: 1) it has the smallest overall $C_p$, 2) the addition of more parameters to the model only increases the $R^2$ in the third decimal place, hardly a substantial improvement. 3) simpler is better.

3. Check the assumptions of this "best" model using all the usual plots (you know what they are by now). Explain in detail whether or not each assumption appears to be substantially violated.

   **Solution:** The qqplot shows that the residuals are approximately normal with some slight deviations from normality at the tails.

   The residual plots show an approximately constant variance, and do not show any obvious outliers. (See Figure 1.)

**For the next three problems use the data from problem 11.15 on page 493 in the text. (`CH11PR15.DAT`)**

4. Plot the data for the two populations on the same graph, using different symbols (`v=`) and lines. Does the relationship between speed and operating cost appear to be the same for the two makes of tire?

   **Solution:** No, the two lines appear to be quite different (see Figure 2). In particular they have different slopes, so that the relationship between cost and speed is different for the two makes of tire.

5. Examine the question of whether or not the two lines are the same. Write a model that allows the two makes of tires to have different intercepts and slopes. Then, perform the general linear test to determine whether the two lines are equal. State the null and alternative hypothesis, the test statistic with degrees of freedom, the $p$-value and your conclusion.

   **Solution:** The model we use is

   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon,$$

   where $X_1 =$ make, $X_2 =$ speed, and $X_3 = X_1 X_2$. This allows the two makes of tire to have different intercepts and slopes.

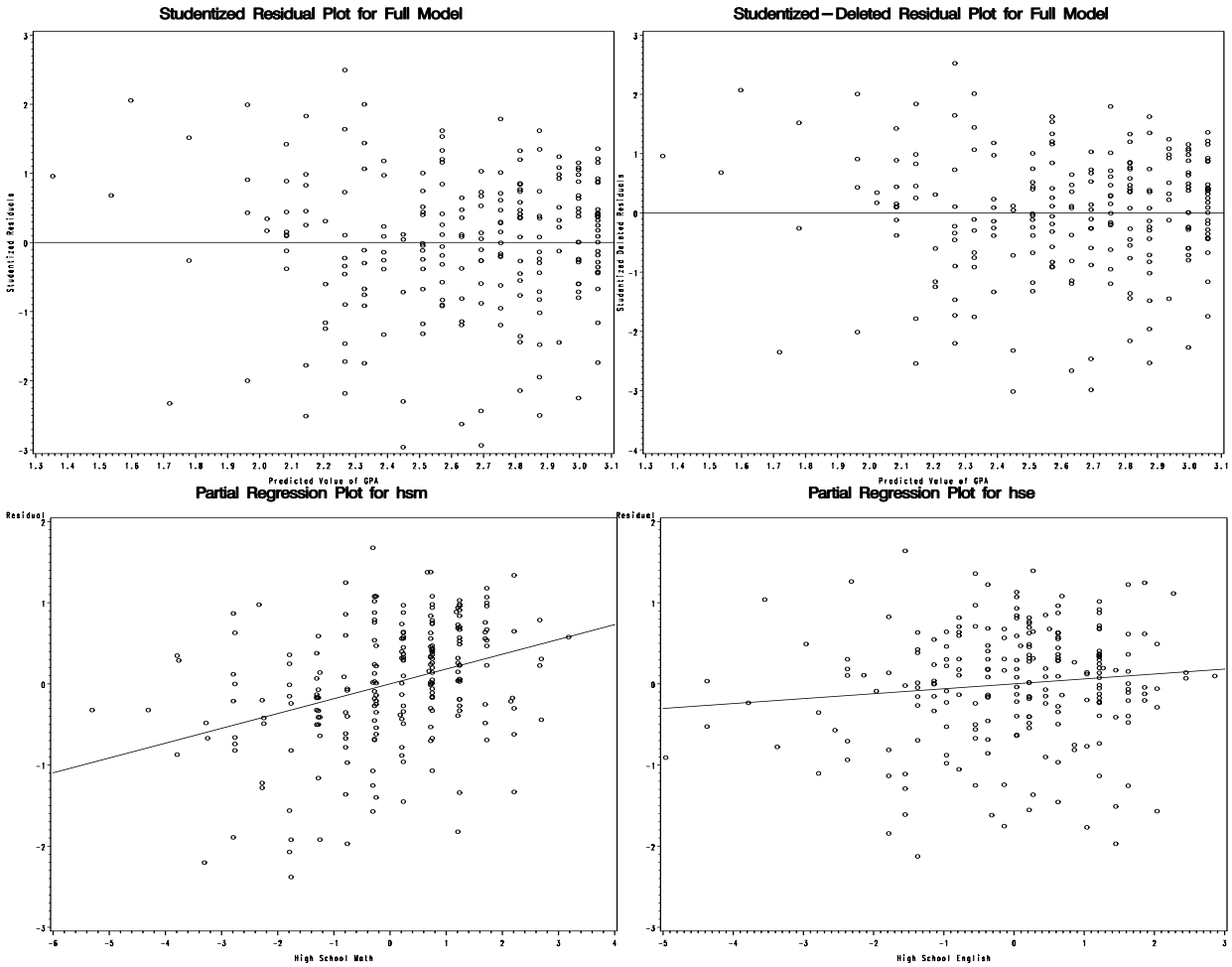


Figure 1: Diagnostic plots for Problem 3

For make A ($make = 0$) the intercept is $\beta_0$ and and the slope is $\beta_2$.
For make B ($make = 1$) the intercept is $\beta_0 + \beta_1$ and the slope is $\beta_2 + \beta_3$.
The null hypothesis is that the two lines have the same slope and intercept.

$$\begin{aligned} \mathrm{H}_0 &: \beta_1 = \beta_3 = 0 \\ \mathrm{H}_a &: \beta_1 \neq 0 \text{ or } \beta_3 \neq 0 \text{ (or both) i.e. } \beta_1 \text{ and } \beta_3 \text{ are not both 0.} \end{aligned}$$

The general linear $F$ ratio for this hypothesis is 12.92 with $(2, 16)$ df, with $p = 0.0005$. We conclude that the two lines are different.

Dependent Variable: cost
Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 3 | 303.86161 | 101.28720 | 84.81 | <.0001 |
| Error | 16 | 19.10789 | 1.19424 | | |
| Corrected Total | 19 | 322.96950 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 1.09281 | R-Square | 0.9408 |

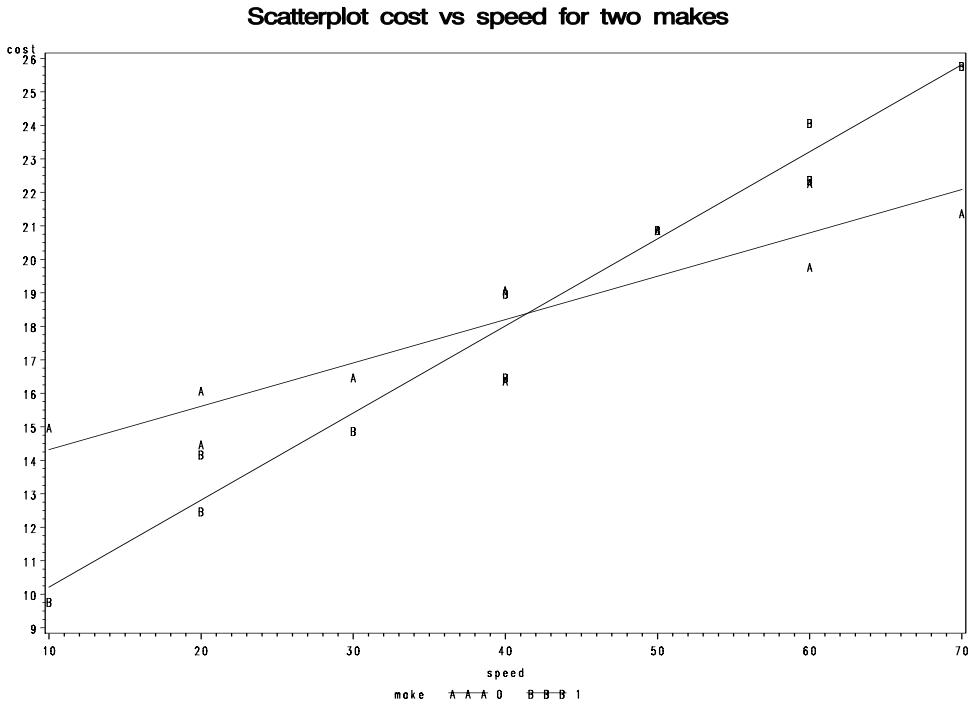


Figure 2: Scatterplot for Problem 4

| | | | |
|---|---|---|---|
| Dependent Mean | 18.10500 | Adj R-Sq | 0.9297 |
| Coeff Var | 6.03598 | | |

Test sameline Results for Dependent Variable cost

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 2 | 15.43053 | 12.92 | 0.0005 |
| Denominator | 16 | 1.19424 | | |

6. Using the model that fits two different lines, give a 95% confidence interval for the difference in slopes. (Hint: what parameter represents the difference between the slopes?)

**Solution:** The difference in slope between the two lines is equal to the parameter $\beta_3$. The 95% CI for $\beta_3$ is $[0.076, 0.185]$.

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | 95% Confidence Limits | |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 13.02222 | 0.80635 | 16.15 | <.0001 | 11.31284 | 14.73161 |
| make | 1 | -5.41222 | 1.14035 | -4.75 | 0.0002 | -7.82966 | -2.99479 |
| speed | 1 | 0.12944 | 0.01821 | 7.11 | <.0001 | 0.09083 | 0.16806 |
| speedmake | 1 | 0.13056 | 0.02576 | 5.07 | 0.0001 | 0.07595 | 0.18516 |