

# Statistics 512: Homework 6 Solutions

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

$$gpa = 0.31101 + 0.003237genderw + 0.11423hsm + 0.03827hss + 0.05103hse + 0.001satm - 0.00041086satv$$

- (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:

$$gpa = 0.34388 + 0.11423hsm + 0.03827hss + 0.05103hse + 0.001satm - 0.00041086satv.$$

- (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:

$$gpa = 0.31101 + 0.11423hsm + 0.03827hss + 0.05103hse + 0.001satm - 0.00041086satv.$$

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		-----Parameter Estimates-----							
Model	C(p)	R-Square	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 8.24 on page 339 in the text. (CH08PR24.DAT)**

4. Plot the data for the two populations on the same graph, using different symbols ( $\mathbf{v=}$ ) and lines. Does the relationship between valuation and selling price 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 price and valuation is different for the two types of lot.

5. Examine the question of whether or not the two lines are the same. Write a model that allows the two lot allocations 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  $X_1 = \text{lot}$ ,  $X_2 = \text{valuation}$ , and  $X_3 = X_1 X_2$ . This allows the two types of lot to have different intercepts and slopes.

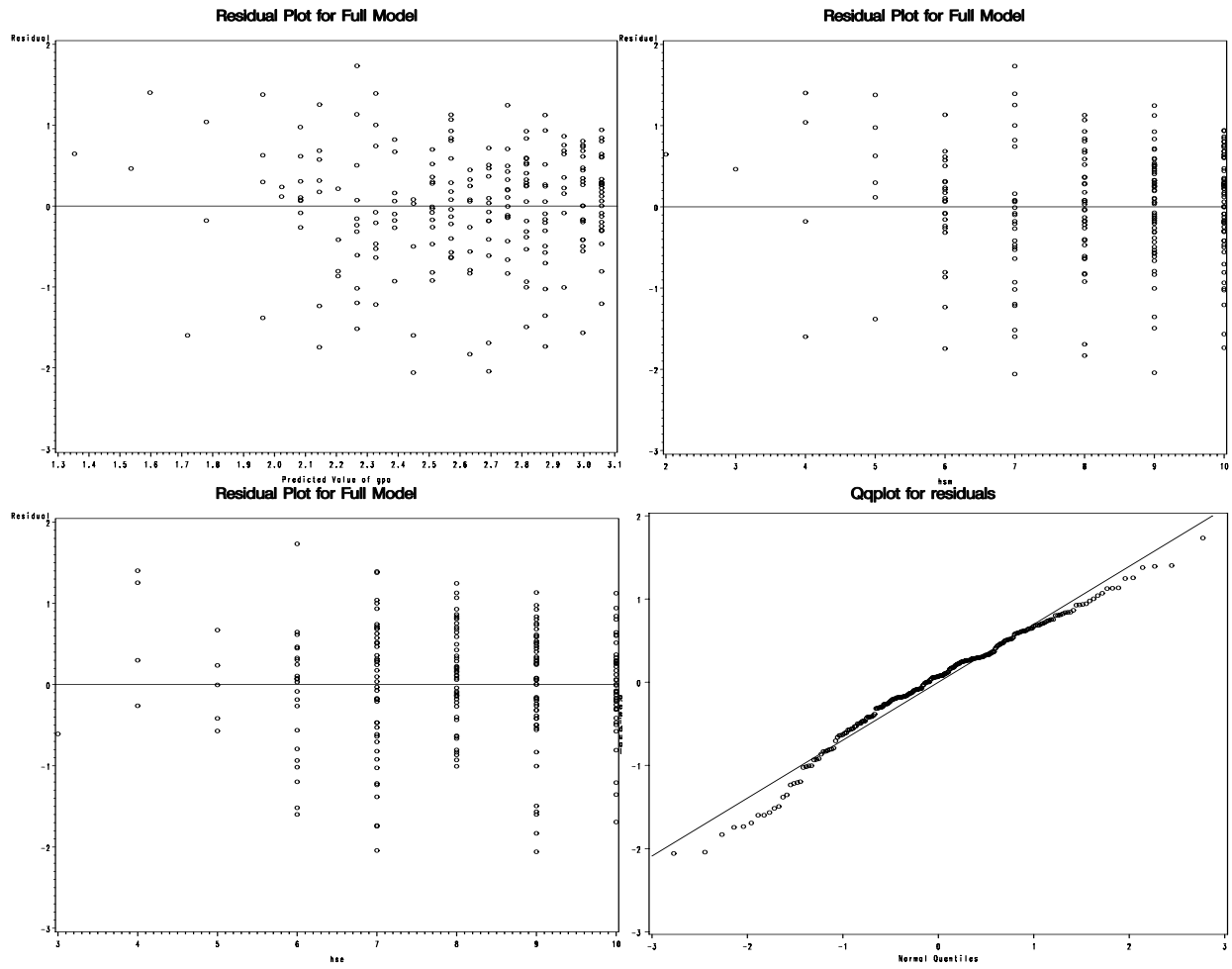


Figure 1: Diagnostic plots for Problem 3

For non-corner lots ( $lot = 0$ ) the intercept is  $\beta_0$  and the slope is  $\beta_2$ .  
For corner lots ( $make = 1$ ) 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.

$$H_0 : \beta_1 = \beta_3 = 0$$

$$H_a : \beta_1 \neq 0 \text{ or } \beta_3 \neq 0 \text{ (or both) i.e. } \beta_1 \text{ and } \beta_3 \text{ are not both 0.}$$

The general linear  $F$  ratio for this hypothesis is 18.68 with (2, 60) df, with  $p = 4.925 \times 10^{-7}$ . We conclude that the two lines are different.

Dependent Variable: price					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4237.05022	1412.35007	93.21	<.0001
Error	60	909.10463	15.15174		
Corrected Total	63	5146.15484			

Test sameline Results for Dependent Variable price

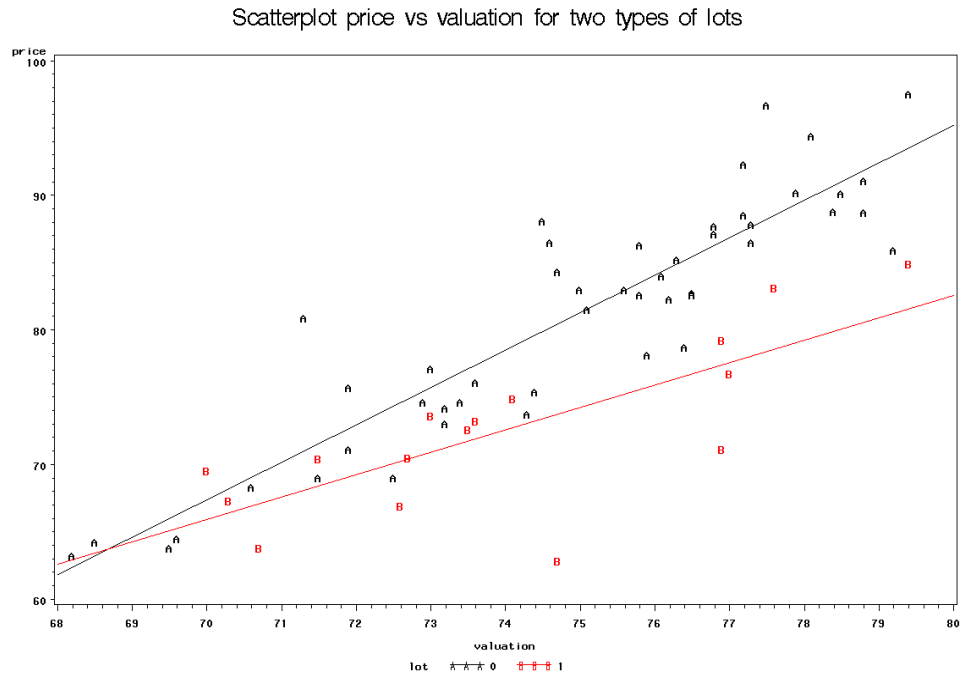


Figure 2: Scatterplot for Problem 4

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	283.07298	18.68	<.0001
Denominator	60	15.15174		

6. Using the model that fits two different lines, give a 94% 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 94% CI for  $\beta_3$  is  $[-1.88491, -0.33006]$ .

		Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	94% Confidence Limits	
Intercept	1	-126.90517	14.72247	-8.62	<.0001	-155.12853	-98.68182
lot	1	76.02153	30.13136	2.52	0.0143	18.25895	133.78412
valuation	1	2.77590	0.19628	14.14	<.0001	2.39962	3.15218
valuationlot	1	-1.10748	0.40554	-2.73	0.0083	-1.88491	-0.33006