Patrick Fitzgerald
ADS 534 Statistical Modeling
HW 5

1. (20 points) To find unusual points in a data set used for simple linear regression $Y_i = \beta_0+\beta_1 X_i+\varepsilon_i$ (i=1,2,...,300), a data analyst examines scatter plot of Y on X as shown in Figure 1. The estimated least square regression line is drawn as well. Sample means $\bar{X} = 24.2$, $\bar{Y} = 49.4$. Classify each of the 5 unusual points on the plot according to type (i.e., Outlier? High leverage point? Influential point?) and justify your answer.

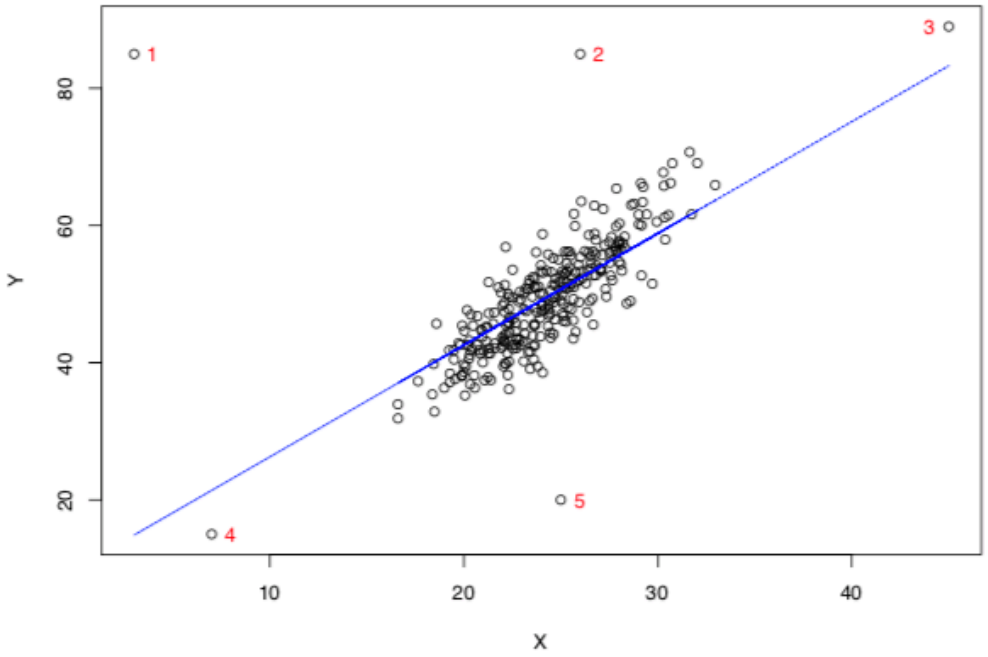


Figure 1: scatter plot of Y on X

(1): this point is an outlier, noted by its large residual and also it is a high leverage point, noted by its isolation in x-space.
(2): this point is an outlier, noted by its large residual
(3): this is a high leverage point, noted by its isolation in x-space
(4): this is a high leverage point, noted by its isolation in x-space
(5): this point is an outlier, noted by its large residual

2. (9 points) Name one or more graphs that can be used to validate each of the following assumptions.
   a. The error terms have constant variance.
      The Studentized residual plot.
   b. The error terms are normally distributed.
      A normal Q-Q plot.

c. There is a linear relationship between the response and predictor variables.

A scatter plot

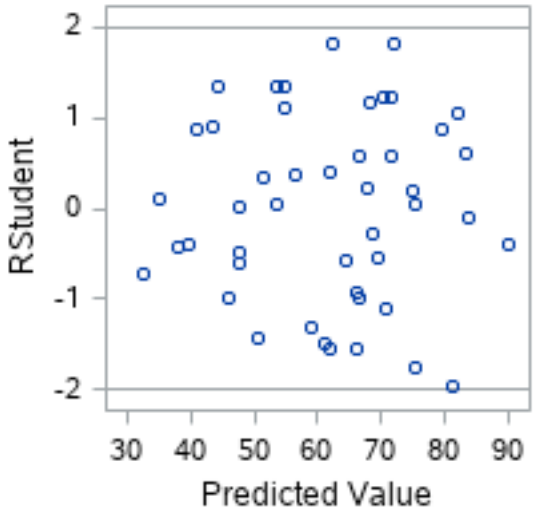
3. (21 points) KNN, problem 10.11, edited. (page 415) A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3, an index).

The administrator randomly selected 46 patients and collected the data in satisfaction.sas7bdat, where larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness and more anxiety.

Fit the regression model

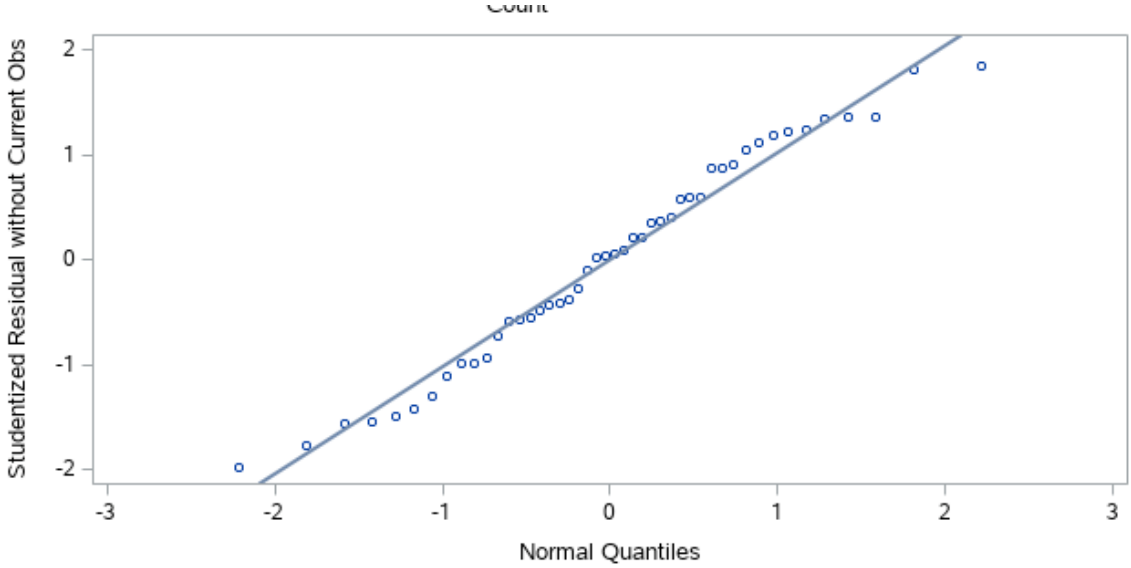
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i,\ i=1,2,\ldots,46.$$

(a) (5 points) Using SAS, obtain the studentized residuals and draw a scatterplot where x-axis is fitted value, and y-axis is the studentized residual. Do there appear to be any outliers?



There are no outliers.

(b) (5 points) Draw a normal probability plot (normal Q-Q plot) using studentized residuals. Interpret your plots and summarize your findings.



The data points follow along the straight line indicating a normal distribution of the Studentized Residuals.

(c) (5 points) Obtain the leverage values of each of the patients using SAS. Is there any high leverage point? (using the rule of thumb discussed in class to make conclusion)

Obs	Y	X1	X2	X3	ypred	r	stud	cookdistance	leverage	sr	dffitsest
1	26	52	62	2.9	32.6597	-6.65970	-0.73311	0.030349	0.18426	-0.72901	-0.34647
2	72	32	46	2.6	66.6051	5.39495	0.59452	0.020194	0.18602	0.58989	0.28200
3	37	47	60	2.4	45.9868	-8.98685	-0.98729	0.053840	0.18096	-0.98698	-0.46393

Leverage values larger than 2(p+1)/n = 2*4/46 = 0.174 will qualify as leverage points (observations 9, 28, and 39 when all listed out).

(d) (6 points) The three largest absolute studentized residuals are for cases 11, 17, and 27. Obtain the Cook's distance values and DFFITS for these cases to assess their influence. What do you conclude?

Obs	Y	X1	X2	X3	ypred	r	stud	cookdistance	leverage	sr	dffitsest
1	89	29	48	2.4	71.8399	17.1601	1.78614	0.07657	0.08759	1.83585	0.56882
2	79	33	56	2.5	62.3904	16.6096	1.75992	0.10513	0.11954	1.80674	0.66574
3	63	25	49	2	81.3524	-18.3524	-1.90944	0.08666	0.08682	-1.97420	-0.60874

We can see from the SAS output that the Cook's distance for these observations are 0.07657, 0.10513, and 0.08666. The DFFITS values for these observations are 0.56882, 0.66574, -0.60874.

See the following R code for the calculations of 20% and 50% for F-distribution with degrees of freedom (n-2+1, n-p) = (4, 42).

```
#### 3(d)

#### find the 20% and 50% percentile for comparison between Cook's
     Distance and F(p+1, n-p-1)

threedone <- qf(0.2, df1=4, df2=42)

threedone

[1] 0.4105709

threedtwo <- qf(0.5, df1=4, df2=42)

threedtwo

[1] 0.8528731
```

The Cook's distance for all 3 observations is less than 20% of the F-distribution thus they have little influence on the fitted values.

Since the sample size, n= 46, is relatively small we use the small data set cutoff

point of 1. Since the absolute values of none of the DFFITS values for the observations are > 1, we conclude that none of the observations are strong influential points.

4. (50 points) A national insurance organization wanted to study the consumption pattern of cigarettes in all 50 states and the District of Columbia. The variables for the study are as follows: The full data set is available in Moodle (cigarette.sas7bdat). The investigators are interested in studying how statewide cigarette consumption is associated with various socioeconomic and demographic variables, and building a parsimonious regression model for predicting the consumption of cigarettes.

   a. (5 points) Build a best linear regression model that predicts the per capita sale of cigarettes in a given state. Perform your analysis using variable selection criterion adjusted R^2. Which variables are included in the final model?

Number in Model	Adjusted R-Square	R-Square	Variables in Model
3	0.2588	0.3032	Age Income Price

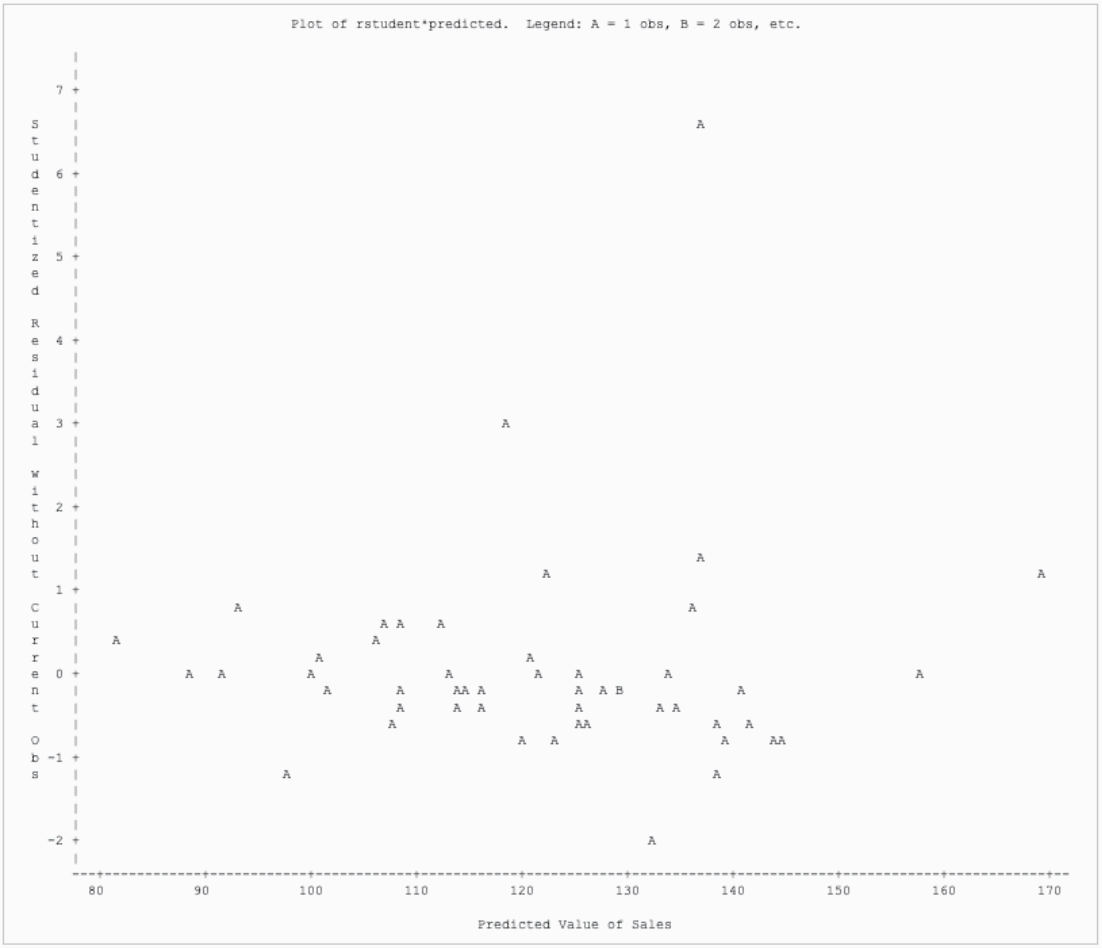
Variable	Definition
Age	Median Age of a person living in the state
HS	Percentage of people > 25 years old who completed high school
Income	Per capita personal income (in dollars)
Black	Percentage of African Americans
Female	Percentage of Females
Price	Weighted average price (in cents) of a pack of cigarettes
Sales	Number of packs of cigarettes sold (per capita basis)

. (b) (5 points) Build a best linear regression model that predicts the per capita sale of cigarettes in a given state. Perform your analysis using variable selection criterion AIC. Which variables are included in the final model?

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	Age	HS	Income	Black	Female	Price	Sales	_IN_	_P_	_EDF_	_RSQ_	_AIC_
1	MODEL1	PARMS	Sales	27.6109	64.248	4.15591	.	0.019281	.	.	-3.39923	-1	3	4	47	0.30324	342.292

Age, Income, and Price are included in the final model with AIC = 342.292

. (c) (5 points) Use studentized residual plot to check whether there is any outlier in response variable Sales?



We see two obvious outliers to be removed from the data set when using a cutoff point of 2 or 2.5.

. (d) (5 points) Delete any outlier you find. Redo part (a). Do you obtain the same final model?

Adjusted R-Square Selection Method

Number of Observations Read	49
Number of Observations Used	49

Number in Model	Adjusted R-Square	R-Square	Variables in Model
5	0.5328	0.5815	Age HS Income Female Price
5	0.5328	0.5814	Age HS Income Black Price

We see two models with equal adjusted R^2 values (0.5328) with variables {Age, HS, Income, Female, and Price} and {Age, HS, Income, Black, and Price}.

. (e) (5 points) Delete any outlier you find. Redo part (b). Do you obtain the same final model?

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	Age	HS	Income	Black	Female	Price	Sales	_IN_	_P_	_EDF_	_RSQ_	_AIC_
1	MODEL1	PARMS	Sales	15.6477	-55.378	.	-0.90960	0.025408	.	4.70455	-3.00901	-1	4	5	44	0.57090	274.258

HS, Income, Female, and Price are included in the final model with AIC = 274.258.

. (f) (5 points) Use the SSE of the final model in part (e) to numerically calculate the AIC (Hint: refer to lecture note 16 slide 16). Compare your AIC with the AIC output from SAS.

It will be shown that the two values are nearly identical.

The following is the R code used to calculate AIC using the given formula:

$$n * \log(SSE / n) + 2 * (p+1)$$

4(f)

manual AIC calculation

```
sse <- 10773

n <- 49

p <- 4

aic <- 49*log(sse/n)+2*(p+1)

aic

[1] 274.2559
```

. (g) (5 points) Interpret the regression coefficients of Age and Black in your final model in part (e).

These variables did not make the final model used.

. (h) (5 points) Delete any outlier you find. Using the forward selection procedure with p-value < 0.10 as the entry criterion, find the final model.

No other variable met the 0.1000 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Price	1	0.1662	0.1662	39.5141	9.37	0.0036
2	Income	2	0.2268	0.3930	18.5270	17.19	0.0001
3	HS	3	0.1362	0.5291	6.7251	13.01	0.0008
4	Female	4	0.0418	0.5709	4.4919	4.28	0.0444

. (i) (5 points) Delete any outlier you find. Using the backward elimination procedure with p-value < 0.10 as the staying criterion, find the final model.

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Black	5	0.0041	0.5815	5.4169	0.42	0.5220
2	Age	4	0.0106	0.5709	4.4919	1.09	0.3024

. (j) (5 points) Delete any outlier you find. Using the stepwise selection procedure with p-value < 0.10 as the entry criterion and p-value < 0.10 as the staying criterion, find the final model.

All variables left in the model are significant at the 0.1000 level.

No other variable met the 0.1000 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Price		1	0.1662	0.1662	39.5141	9.37	0.0036
2	Income		2	0.2268	0.3930	18.5270	17.19	0.0001
3	HS		3	0.1362	0.5291	6.7251	13.01	0.0008
4	Female		4	0.0418	0.5709	4.4919	4.28	0.0444