Statistics 512: Homework 1

Solutions

1. A regression analysis relating test scores (Y) to training hours (X) produced the following fitted question: $\hat{y} = 25 - 0.5x$.

   (a) What is the fitted value of the response variable corresponding to $x = 7$?

   Solution: The fitted value at $x = 7$ is

   $$\hat{y} = 25 - (0.5)(7) = 25 - 3.5 = 21.5.$$

   (b) What is the residual corresponding to the data point with $x = 3$ and $y = 30$?

   Solution: The fitted value corresponding to the data point with $x = 3$ is

   $$\hat{y} = 25 - (0.5)(3) = 25 - 1.5 = 23.5.$$

   The residual corresponding to the data point with $x = 3$ and $y = 30$ is thus

   $$e_i = 30 - 23.5 = 6.5.$$

   (c) If x increases 3 units, how does $\hat{y}$ change?

   Solution: For increase of 1 in x, $\hat{y}$ changes by the slope. Therefore, if x increase by 3 units, $\hat{y}$ will decrease by 3 times the slope, i.e. by $(3)(-0.5) = -1.5$.

   (d) An additional test score is to obtained for a new observation at $x = 6$. Would the test score for the new observation necessarily be 22? Explain.

   Solution: Not necessarily. The new observation is a random variable from a normal distribution with estimated mean 22 . So you would not likely see the observation 22 .

   (e) The error sums of squares (SSE) for this model was found to be 7. If there were $n = 16$ observations, provide the best estimate for σ^2.

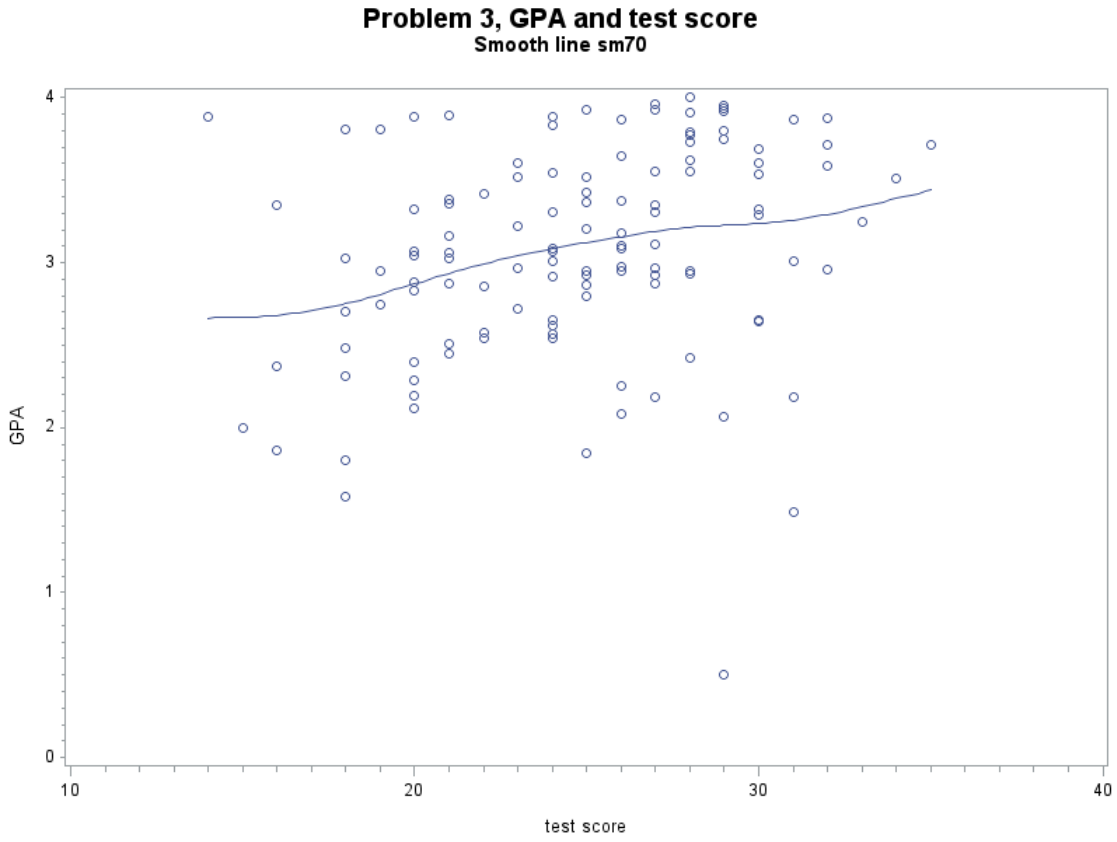
   Solution: The estimate of σ^2 is given by

   $$s^2 = MSE = \frac{SSE}{df_E} = \frac{SSE}{n-2} = \frac{7}{14} = 0.5.$$

2. Explain the different between the following two equations:

   $$\begin{aligned} \hat{Y} &= b_0 + b_1 X \\ Y &= \beta_0 + \beta_1 X + \epsilon. \end{aligned}$$

   Solution: The first equation is the fitted regression line which describes the linear relationship between the *mean* of the response variable (fitted value) and the explanatory variable X. The second equation is the linear model which describes the relationship between the observed (X, Y) pairs. Not all the pairs will fall directly on a line as they will in the first equation, as indicated by the error term. In the first equation, the values b_0 and b_1 are known values obtained from the data, while in the second equation, the parameters β_1 and β_0 are unknown.



3. For this problem, use the "grade point average" data described in KNNL Problem #1.19. The data are on the disk that accompanies the text and can also be found on the class web site (`CH01PR19.DAT`). Make sure you understand which column is X and which is Y and read in the data accordingly. See Topic 1 or `knnl054.sas` for an example of how to read in a data file.

   (a) Plot the data using `proc gplot`. Include a smoothed function on the plot by preceding the `plot` statement with "`SYMBOL1 v = square i = smNN`" where `NN` is a number between 1 and 99. Note that larger numbers cause greater smoothing. For this homework please use the number 70. Make sure to indicate the smoothing number in the title of the plot. Is the relationship approximately linear?

   Solution: The relationship looks approximately linear. (See graph).

   (b) Run a linear regression to predict GPA based on the entrance exam. Give the complete ANOVA table for this regression.

   Solution: The included table shows that the linear model does indeed explain a significant portion of the variability in the response.

Problem 3, GPA and test score
Smooth line sm70

The REG Procedure
Model: MODEL1
Dependent Variable: gpa

Number of Observations Read	120
Number of Observations Used	120

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	3.58785	3.58785	9.24	0.0029
Error	118	45.81761	0.38828		
Corrected Total	119	49.40545			

Root MSE	0.62313	R-Square	0.0726
Dependent Mean	3.07405	Adj R-Sq	0.0648
Coeff Var	20.27049		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	94% Confidence Limits	
Intercept	1	2.11405	0.32089	6.59	<.0001	1.50465	2.72344
score	1	0.03883	0.01277	3.04	0.0029	0.01457	0.06308

Analysis of Variance

Source	DF	SS	MS	F	p
Model	1	3.588	3.588	9.24	0.0029
Error	118	45.818	0.38828		
Corrected Total	119	49.405			

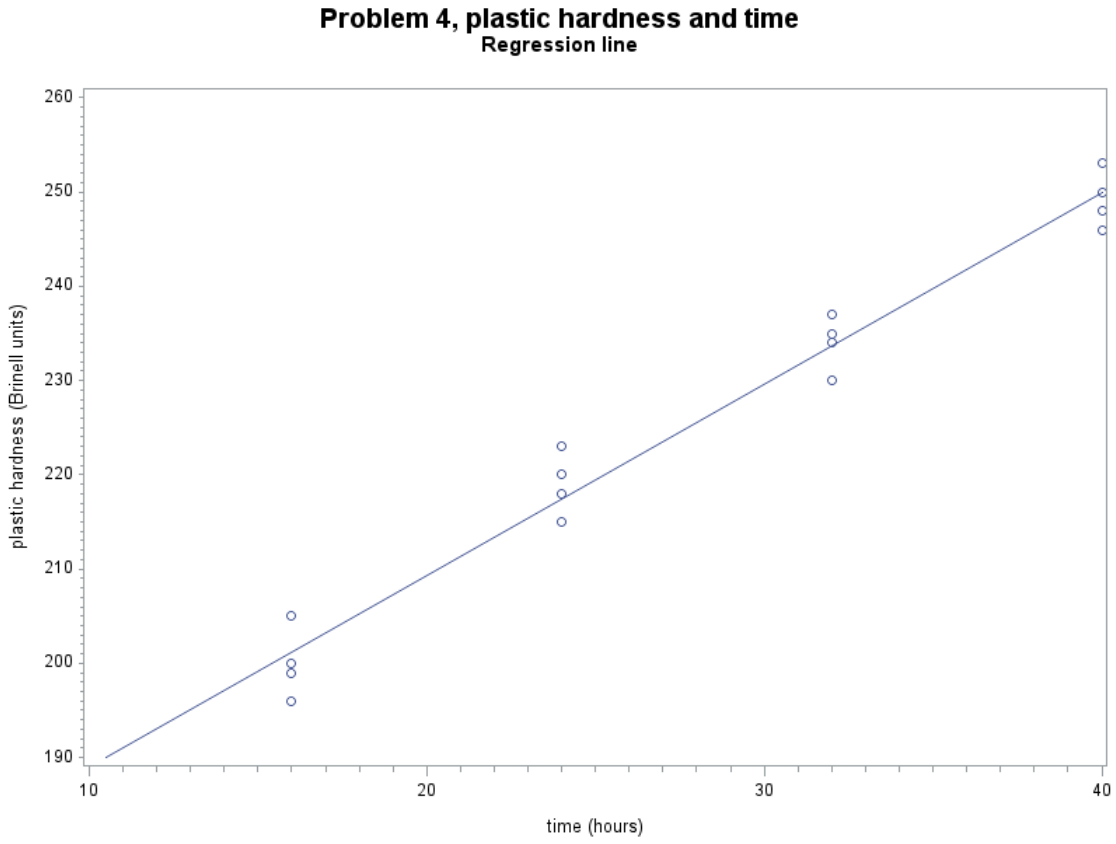
(c) Give a point estimate and a 94% confidence interval for the slope and intercept and interpret each of these in words.

Solution: Based on the SAS output (see file), the point estimate for the slope is $b_1 = 0.0388$. This is our best least-squares estimate for the slope of the regression line. The 94% CI for β_1 is the interval $[0.01457, 0.06308]$. We are 94% confident that β_1 is in the interval.

In addition, the point estimate for β_0, the intercept, is 2.114, estimated via least squares estimation. The interval $[1.50465, 2.72344]$ is the 94% confidence interval for this parameter. We are 94% confident that the true value β_0 is in this interval.

(d) Would it be reasonable to consider inference on the intercept for this problem? Please provide justification for your answer.

Solution: It is doubtful that inference on β_0 would be reasonable in this problem. The interpretation of β_0, if any, would be the mean GPA obtained by someone with a score of 0 on the test. However, none of the observations in our



sample had test scores below 3.9 . Therefore, it's difficult to say that the linear relationship would continue for lower test scores; in fact, we can be certain that it would not, since a negative GPA is not possible.

4. For this problem, use the "plastic hardness" data described in the text with problem 1.22 on page 39. (`CH01PR22.DAT`) Make sure you understand which column is X and which is Y and read in the data accordingly.

   (a) Plot the data using `PROC GPLOT`. Include a regression line on the plot (`i = rl`). Is the relationship approximately linear?

   Solution: The graph seems to indicate that the relationship is quite close to linear.

   (b) Run the linear regression to predict hardness from time. Give

   i. the linear model used in this problem

   Solution: The linear model is $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$.

   ii. the estimated regression equation.

   Solution: The estimated regression equation is $\hat{Y} = 168.6 + 2.034X$.

   (c) Describe the results of the significance test for the slope. Give the hypotheses being tested, the test statistic with degrees of freedom, the p-value, and your conclusion in sentence form.

   Solution: The null hypothesis is $\mathrm{H}_0 : \beta_1 = 0$, and the alternative is $\mathrm{H}_A : \beta_1 \neq 0$. The observed test statistics is $t = 22.51$, with degrees of freedom $n - 2 = 14$

Problem 4, plastic hardness and time
Regression line

The REG Procedure
Model: MODEL1
Dependent Variable: hard

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5297.51250	5297.51250	506.51	<.0001
Error	14	146.42500	10.45893		
Corrected Total	15	5443.93750			

Root MSE	3.23403	R-Square	0.9731
Dependent Mean	225.56250	Adj R-Sq	0.9712
Coeff Var	1.43376		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	168.60000	2.65702	63.45	<.0001
time	1	2.03438	0.09039	22.51	<.0001

and p-value 1.08×10^{-12}. We reject H_0 and conclude that the slope is non-zero. So there is a significant linear relationship between hardness and time.

(d) Explain why or why not inference on the intercept is reasonable (i.e. of interest) in this problem.

Solution: Since we have no data near $X = 0$, only for X between 16 and 40, we cannot be certain that the linearity of the relationship continues to hold in that region. The interpretation of the intercept would be the hardness of the plastic immediately after it is molded. The hardness at that time might behave quite differently then than after 16 hours. For example, there could be an initial rapid hardening of the plastic as it cools, followed by a more gradual response at room temperature. So while the hardness at time 0 might be of interest, we cannot use the intercept to make inference about it based on these data.

5. An investigative study collected 40 observations from the Wabash river at random locations near Lafayette. Each observation consisted of a measure of water pH (X) and fish count (Y). The researchers are interested in how the acidity of the water affects the number of fish. Complete the following ANOVA table for the regression analysis (the p-value need not be exact). State the null and alternative hypotheses for the F-test as well as your conclusion in sentence form.

Solution: The completed ANOVA table looks like

Source	degrees of freedom	Sum of Squares	Mean Square	F-value	Pr $> F$
Model	**1**	55.30	**55.30**	**445.97**	$<$ **0.0001**
Error	**38**	**4.70**	**0.124**		
Corrected Total	**39**	60.00			

The hypotheses are

$$\begin{aligned} \mathrm{H}_0: &\quad \beta_1 = 0 \\ \mathrm{H}_A: &\quad \beta_1 \neq 0. \end{aligned}$$

We reject the null hypothesis and conclude that there is a significant linear relationship between water pH and fish count.