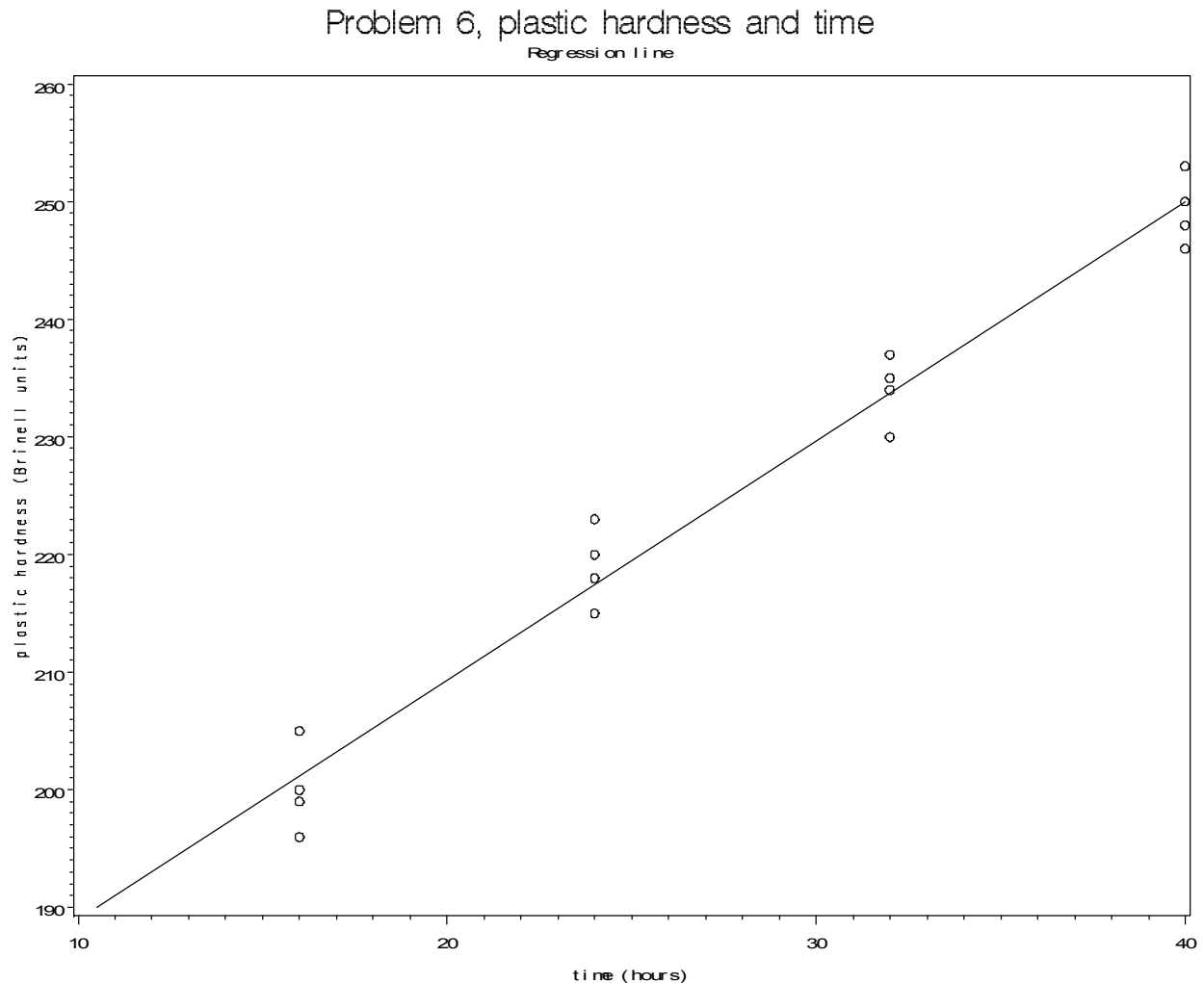


1. For this problem use the “plastic hardness” data described in the text with problem 1.22 on page 36. (CH01PR22.DAT). Make sure you understand which column is X and which is Y and read in the data accordingly.

(a) Plot the data using PROC GPLOT. Include a regression line on the plot (i=r1). Is the relationship approximately linear?



The relationship looks quite linear.

(b) Run the linear regression to predict hardness from time. Give (see problem 2)

i. the linear model used in this problem

The linear model is $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.

ii. the estimated regression equation.

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

The estimated regression equation is $\hat{Y} = 168.6 + 2.03438X$.

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

Error	14	146.42500	10.45893
-------	----	-----------	----------

The null hypothesis is $H_0: \beta_1 = 0$, and the alternative is $H_A: \beta_1 \neq 0$. The observed test statistic is $t = 22.51$, with degrees of freedom $n-2 = 14$ and P-value $P < 0.0001$. We reject H_0 and conclude that the slope is non-zero, so that 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.

Because 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 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.

SAS Code

```

title1 'Problem 5, GPA and test score';
data prob5;
    infile 'ch01pr19.dat';
    input gpa score;
proc print data=prob5;
    * Sort the data by the X variable to avoid a scribbly line;
proc sort data=prob5;
    by score;
title2 'Smooth line sm70';
axis1 label=('test score');
axis2 label=(angle=90 'GPA');
symbol1 v = circle i=sm70;
proc gplot data=prob5;
    plot gpa*score/haxis=axis1 vaxis=axis2;
proc reg data=prob5;
    model gpa=score/clb;
run;
title1 'Problem 6, plastic hardness and time';
title2 'Regression line';
data prob6;
    infile 'ch01pr22.dat';
    input hard time;
proc print data=prob6;
symbol1 v=circle i=rl;
axis1 label=('time (hours)');
axis2 label=(angle=90'plastic hardness (Brinell units)' );
proc gplot data=prob6;
    plot hard*time/haxis=axis1 vaxis=axis2;

```

```
proc reg data=prob6;
    model hard=time;

run;
quit;
```

- e) Give an estimate of the **mean** hardness that you would expect after 36 and 43 hours; and a 95% confidence interval for each estimate. Which confidence interval is wider and why is it wider?

Output Statistics

Obs	time	Dep Var hard	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
17	36	.	241.8375	1.0847	239.5110	244.1640	234.5214	249.1536
18	43	.	256.0781	1.5787	252.6922	259.4640	248.3595	263.7967

X = 36: $\hat{Y} = 241.8375$, CI for the mean is [239.5110, 244.1640]

X = 43: $\hat{Y} = 256.0781$, CI for the mean is [252.6922, 259.4640]

The confidence interval for X = 43 is wider because 43 is farther away from the mean \bar{X} than is 36, and so the standard error for the prediction is larger.

- f) Give a prediction for the hardness that you would expect for an **individual** piece of plastic after 43 hours; and a 95% prediction interval for this quantity.

X = 43: $\hat{Y} = 256.0781$, CI for an individual observation is [248.3595, 263.7967]

2. You fit the simple linear model to a data set and obtain estimates $b_0 = 1$, $b_1 = 3$, and $s = 3.0$

- a) With $n = 18$ cases, there are 16 degrees of freedom for the t distribution. For a 95% CI, you'd use $t_c = 2.12$. Thus the confidence interval would be $3 \pm 2.12(1) = (0.88, 5.12)$

- b) Since zero is not in the confidence interval, we'd reject the hypothesis test that $H_0: \beta_1 = 0$ at the $\alpha = 0.05$ level. Since the slope is something other than zero, this suggests that X is useful in predicting Y.

c) The predicted value is found in the middle of the interval. In this case, it is 16. To form the 95% prediction interval, one needs to determine the variance for the predicted value. Using the form of the confidence interval, we can compute the standard error of the mean at $X = 5$. Again with 16 df, $t_c = 2.12$. Since the interval is 16 ± 2.12 , this implies the standard error of the mean at $X = 5$ is 1. The variance of the predicted value is then $1^2 + 3^2 = 10$ where $3^2 = 9$ is the MSE. Thus the prediction interval is $16 \pm 2.12\sqrt{10} = (9.30, 22.70)$.

(KNNL) 2.1. a. Yes, $\alpha = .05$

(KNNL) 2.26.

a.

Source	SS	df	MS
Regression	5,297.5125	1	5,297.5125
Error	146.4250	14	10.4589
Total	5,443.9375	15	

c.

$i:$	1	2	3	4	5	6
$Y_i - \hat{Y}_i:$	-2.150	3.850	-5.150	-1.150	.575	2.575
$\hat{Y}_i - \bar{Y}:$	-24.4125	-24.4125	-24.4125	-24.4125	-8.1375	-8.1375
$i:$	7	8	9	10	11	12
$Y_i - \hat{Y}_i:$	-2.425	5.575	3.300	.300	1.300	-3.700
$\hat{Y}_i - \bar{Y}:$	-8.1375	-8.1375	8.1375	8.1375	8.1375	8.1375
$i:$	13	14	15	16		
$Y_i - \hat{Y}_i:$.025	-1.975	3.025	-3.975		
$\hat{Y}_i - \bar{Y}:$	24.4125	24.4125	24.4125	24.4125		