STAT 512: Homework 1
Name: Lei Nie

1. *A regression analysis relating scores in an aptitude test after training (Y) to training hours (X) produced the following fitted question:* $\hat{y} = 25 + 1.4x$.

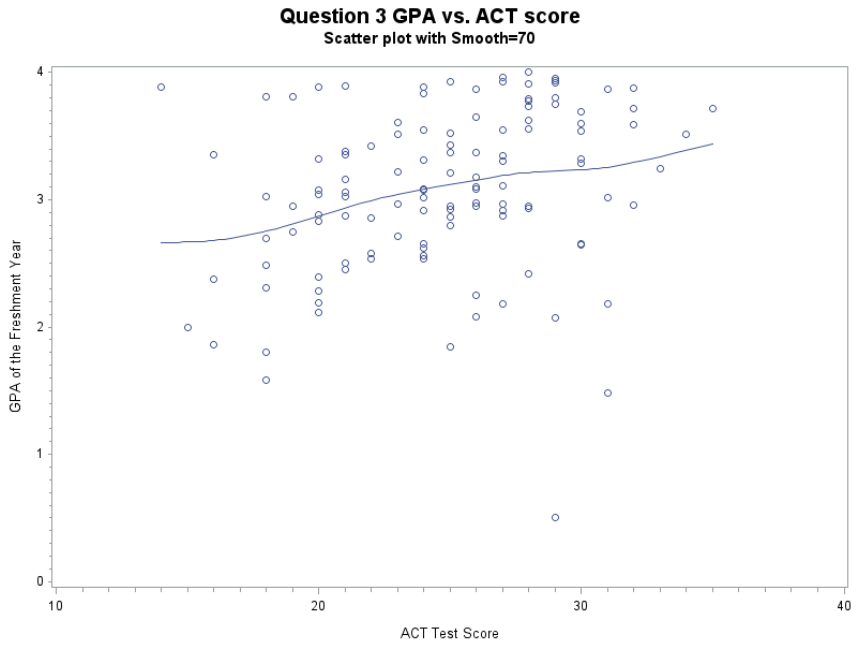
   (a) When $x = 5$, $\hat{y} = 25 + 7 = 32$.
   (b) When $x = 8$, $\hat{y} = 36.2$. $\epsilon = y - \hat{y} = 0$. The point is on the line.
   (c) When x increases 2 units, $\hat{y}$ increases 2.8 units.
   (d) Not necessary. The test score of the new observation is a random variable with mean 34.8. Statistically speaking, the probability for y to be equal to a certain value is very small.
   (e) $\widehat{\sigma^2} = \frac{SSE}{n-2} = \frac{9}{18} = 0.5$.

2. *Explain the difference between the following two equations: Also, describe the distribution of* Y_i *and* ϵ_i. *What is the meaning of* β_0 *and* β_1*?*

   (a) $Y_i = b_o + b_1 X_i$
   The equation (a) is to calculate the mean of the fitted value using the explanatory variable X. The b_0 and b_1 in the equation are the estimation of β_0 and β_1 in the equation (b). They are known.

   (b) $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
   The equation (b) is the linear model describing the relationship between the observation value X and Y. The β_0 and β_1 are unknown parameters.

3. *Grade Point Average*
   (a) Draw the scatter plot with a smoothing curve.



The relationship seems approximately linear.

(b) Run a linear regression model.

ANOVA table:

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		

(c) Give a point estimate and 90% confidence interval for the slope and intercept.

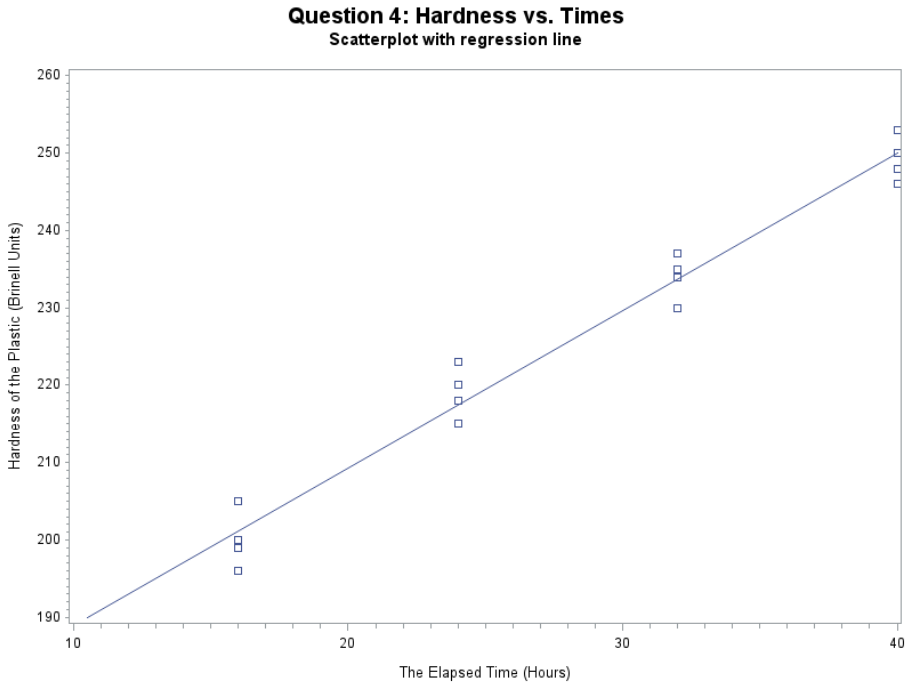
Parameter Estimates							
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**90% Confidence Limits**	
Intercept	**1**	2.11405	0.32089	6.59	<.0001	1.58205	2.64605
ACT	**1**	0.03883	0.01277	3.04	0.0029	0.01765	0.06000

Intercept: The estimated slope is 2.11405. We are 90% confident that the true value of the slope is in the range of (1.58205, 2.64605)
Slope: The estimated slope is 0.03883. We are 90% confident that the true value of the slope is in the range of (0.01765, 0.06000)

(d) Would it be reasonable to consider inference on the intercept for this problem?
No, because zero is not in the range of the ACT.

4. *Plastic Hardness*

(a) Plot the data using PROC GPLOT with a regression line.



The relationship seems approximately linear.

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

i. The linear model used in this problem

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

ii. The estimated regression equation.

$$\hat{Y} = b_0 + b_1 X = 168.6 + 2.0344X$$

(c) Describe the results of the significance test for the slope.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	**1**	168.60000	2.65702	63.45	<.0001	162.90125	174.29875
time	**1**	2.03438	0.09039	22.51	<.0001	1.84050	2.22825

$$H_0: \beta_1 = 0 \; vs. \; H_a: \beta_1 \neq 0$$

Since the test statistic $t = 22.51$ with the degrees of freedom 14, $P - value < \alpha$ $(\alpha = 0.05)$. With 95% confidence, we reject the null hypothesis and conclude that there is a 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.

The inference on the intercept is not reasonable because zero is not in the range of the ACT. The intercept represents the mean hardness of the plastic when it was just

removed from the mold. The hardness might behave completely differently from that after certain hours.

5. *Fish count under different water PH value. Complete the following ANOVA table for the regression analysis. State the null and alternative hypotheses for the F-test as well as your conclusion in sentence form.*

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	MSM* df_M	MSM=46.7	MSM/MSE	
Error	n-2	SSE=83.7	SSE/df_E		
Corrected Total	n-1	SSM+SSE			

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	46.7	46.7	82.5759	<0.0001
Error	148	83.7	0.5655		
Corrected Total	149	130.4			

$$H_0: \beta_1 = 0 \; vs. \; H_a: \beta_1 \neq 0$$

Since the test statistic $F = 82.5759$ with the degrees of freedom 1 and 148, $P - value = 5.5511\text{E} - 16 < \alpha \; (\alpha = 0.05)$. With 95% confidence, we reject the null hypothesis and conclude that there is a linear relationship between pH value and fish number in the Wabash River.

Appendix: SAS code

```
*Question 3;
data grade;
input GPA ACT @@;
datalines;
 3.897    21
……
2.948    28
;
proc print data=grade;run;

*Question 3(a);
proc sort data=grade;by ACT;
symbol v = circle i = sm70;
title1 'Question 3 GPA vs. ACT score';
title2 'Scatter plot with Smooth=70';
axis1 label=('ACT Test Score');
axis2 label=(angle=90 'GPA of the Freshment Year');
proc gplot data=grade;
plot GPA*ACT/haxis=axis1 vaxis=axis2;
run;

*Question 3(b)(c);
proc reg data=grade;
model GPA=ACT/clb alpha=0.1;
run;

*Question 4;
data plastic;
input hardness time;
datalines;
  199.0   16.0
……
  246.0   40.0
;
proc print data=plastic;run;

*Question 4(a);
proc sort data=plastic; by time;
symbol1 v=square i=rl;
title1 'Question 4: Hardness vs. Times';
title2 'Scatterplot with regression line';
axis1 label=('The Elapsed Time (Hours)');
axis2 label=(angle=90 'Hardness of the Plastic (Brinell Units)');
proc gplot data=plastic;
plot hardness*time/ haxis=axis1 vaxis=axis2;
run;

*Question 4(b);
proc reg data=plastic;
model hardness=time/clb;run;

*Question 4(c);
data a4;pvalue=1-probt(22.51,14);
proc print data=a4;run;

*Question 5;
data a5;pvalue=1-probf(82.575945,1,148);run;
proc print data=a5;run;
```