STAT 2131:
Applied Statistical Methods I
HW #3
Due Tuesday 11:00am, October 12th

1. The army would like to explore the relationship between smoking and lung capacity. A study was conducted in which $n = 32$ new recruits reported the number of cigarettes they smoked within the last 7 days before having their lung capacity measured. The data set `smoking` contains two variables: `cigs` is the number of cigarettes smoked in the past week and `LC` is lung capacity in mL.

   The following SAS code was run to obtain the ensuing output:

```
proc reg data=smoking;
model LC=cigs / clb;
output out=CigOut P=yhat STDP=s STDI=spred;
run;

proc print data=CigOut; run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: LC

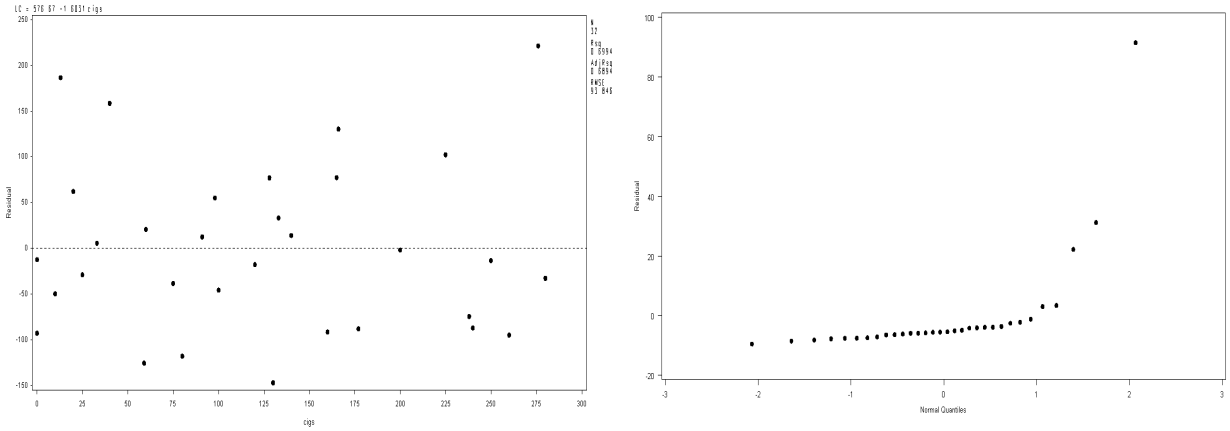
Number of Observations Read 32
Number of Observations Used 32

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	544551	544551	54.66	<.0001
Error	30	298890	9962.99336		
Corrected Total	31	843440			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	576.67459	29.12407	19.80	<.0001	517.1953	636.1538
cigs	1	-1.60309	0.19188	-8.35	<.0001	-1.99497	-1.21122



(a) Write down the assumed model. Clearly state any assumptions and define all notation.

(b) Assume that the model in part (a) is appropriate. Provide a point estimate and a 95% confidence interval for the change in expected lung capacity if a recruit increases his smoking by 1 pack of cigarettes per week. One pack consists of 20 cigarettes.

(c) A scatter plot of the residuals from the fitted regression model versus the number of cigarettes smoked and a normal residual qq-plot were created. These are are displayed above: the residual scatter plot is on the left and the qq-plot is on the right. Answer the following three questions based on the information contained in theses plots.

(i) What assumption of the model in part (a) do you feel has been violated? Why?

(ii) Are you concerned that the point estimate reported in part (b) is biased? Why?

(iii) Are you concerned that the confidence interval reported in part (b) is inaccurate? Why?

2. You are working for a company that developed a new innovation for the insurance industry. You are interested in knowing how the amount of waiting time (Y_i in months) it takes an insurance firm to adopt your innovation is related to the size of the firm (in million dollars) and the type of the firm (mutual or stock). The data is in the file "stock.txt". Fit linear regression models to answer the following questions.

(a) Fit a model with interaction, i.e., "y $\sim$ size + type + size $\times$ type", and answer the following three questions.

(i) What is the estimated change in the waiting time for a 1 million dollar increase in size for a mutual fund company.

(ii) What is the estimated change in the waiting time for a 1 million dollar increase in size for a stock company.

(iii) What is the estimated difference in the waiting time between a stock company of size X_1 and a mutual fund company of size X_1.

(b) Is the interaction term statistically significant at 0.05 level? Conduct a hypothesis test.

(c) Re-fit the model without interaction and answer the three questions in part (a).