06.11.2015 (dd.mm.yy)一模一样的题目：

3，5，6，7，8，16，17，20，21，24，29，30，32，34，41(填空），43，44，45，47，48，49，51，52，53，55，56，58，60，65

题目改动：

9，model A has higher accuracy

15, ninteen times higher

22，改动选项，有一个是only XL and 2XL，有一个是XL and 2XL only

23，改动数字

25，改动图片，右边的竖线在阴影外面

28，missing改成了redundant，答案varclus

33，including改成了eliminating

35，问如何改进，选择log transformation

36，Spearman corrleation，选择outs

39, excluded，选decrease

50，collinearity改成了influencial factor,选cooksd

57，concordant改成了discordant

61，改成better model，选adjusted R square

新加题型：

1， learn how to read interaction graph, decide what indicates interaction between income and high value(low, medium, high)?

A.all intercepts between income and low, high value

B.no intercept

C.?

D.intercepts between income and high value

2, reading ANOVA output for multiple variables.

based on P-value<0.05, what can you tell?

Answer:at least two of the means are different from each other

3, ANOVA output.

two outputs, one from the model, P-value<0.05, one about S, M, L and three P-values. Question: determine if S,M,L are significant.

4, Which of the following is improper use of PROC LOGISTIC?

A.predict loan default.....

B.predict buy houses.....

C.predict usage of internet to do.....

D.rank the liklihood of default on loans.....

5, Accuracy, Error rate calculation

6, Y=a+b*X1+c*X2

now we know X1 is 30, X2 is unkown, what is the predicted value of Y?

Answer: unpredicatable

7, There is large difference between training and test data, what does this indicate?

Answer: Overfitting

8, Given output with VIFs, what can you tell about collinearity?

Answer: There is not significant collinearity becaue VIF<10

9, Which of the following is correct for LOGISTIC with Male/Female, High/Low?

Answer: Codes with Param=, Ref=

10，Which of the following obey the hierarchy=single principle?

A. X Y

B. X*Y

C. X X*Y

D. X Y X*Y

Answer: A,D

11，Which of the following generate the ROC curve? A.

```
proc reg;
     model count = degree belief;
     roc belief;
run;
```

B. proc roc;

?

C. proc genmod order=data;

```
    model count = degree belief / dist=poi;
run;
```

D. proc logistic;

```
    model count = degree belief;
     roc belief;
run;
```

11.06.2014(dd.mm.yy) 考试机经以及题库题目分析
Question 1、2 没有考

Question 3

考了原题

Question 5

考了原题

Question 6

没有考

Question 7

考了原题

并且新加了一题，考 gains chart 的 reference line 意义

Question 8

没有考，不过挺有意义

分析如下：

The following SAS program is equal to:

```
Data new;
    Set new;
    Solicit=PE>0.5;
Run;
Proc means data=new mean;
     Var solicit;
```

Where TG=1;

Run;

因此，在第二个程序中输入的 sub-dataset 仅仅包含了 TG=1 的情况（即 event 实际发生的部分）。求出的均值相当于 sensitivity。Choose B

Question 9

考了变体，主要学会分析即可

比较模型时，主要看 validation 部分的 accuracy。仅仅看 training 部分不够，会有 overfitting 情况出现。

Therefore， choose D

Question 10：

考了原题

注意，200 为 profit，不是 revenue。

		Solicit	
Purch		0	1
	0	0	-10
	1	0	200

Profit=（P_R>0.05）*Purch*200+（-10）*（1-purch）*（P_R>0.05）

Therefore， choose A

Question 11：

没有考

Honest assessment 一定有 training data 和 validation data， testing 可以没有

Training data：model generation /build a model

Validation data：model comparison and assessment

Testing data：provide an unbiased measure of assessment for the final model.

Choose D

Question 12：

没有考

ROC curve 下的面积越大，模型越好

Choose B

Question 13：

考了变体

注意此处 500 为 revenue， 我们要先转换为 profit。500-50=450 profit

Choose C

Question 14：

原题

Sensitivity and specificity are not affected by oversampling.

书上原话

Sensitivity and specificity, however, are not affected by separate sampling because they do not depend on the proportion of each class in the sample.

Choose D

Question 15

考了变体，数据不同了！！！要自己学会算啊

Priorevent=the probability of events in population=0.05

P_1 范围从此求出：

		Solicit	
Respond		0	1
	0	0	-1
	1	0	9

9*pi+（-1）*pi>0 ------pi>0.1

Choose B

Question 16

一模一样原题

Question 17

考了变体，填空题，不过数据不一样，要自己学会算

Sensitivity=true positives/total actual positives

=25/(23+25)

=25/28

除此以外，还考了另外一题比较类似

要求求 Accuracy 和 error rate。记住公式哦~~

Question 18，19 20 没有考

Question 21 考了填空，要求填入 hovtest

Question 22 考了原题

Question 23.

原题，不过自己填空算

R_squared=1-SSE/SST=SSR/SST=15716/20761=0.76

Question 24 没有考，不过考了变体，看 GLM 生成的图，找明显的 assumption violation

那些图中 QQ plot 很明显不是斜对角线，以及 histogram 明显不是正态分布，所以我选了 the assumption of normality is violated

Question 25

考了变体，大家自己要学会读图

第一， 注意 control=S，就是说用 small size 与其他组（medium，large）比较，small 为 reference 组

第二， 看阴影部分：若竖线在阴影部分之内，说明 not significantly different

若竖线在阴影部分之外，说明 significantly different

Choose C

Question 26， 考了原题

Question 27， 考了原题

Question 28 ,考了变体，问 variable cluster 用哪个，选 varclus

Question 29， 考了原题

Question 30，考了原题，我记得选 A。。。

Question 31, 考了原题

Question32，考了原题

Question 33：

这题没有考原题，考了变体：

考试题目：Excluding redundant input variables in a regression model can:

Stabilize parameter estimates and decrease the risk of overfitting.

Question 34， 考了原题

Question 35， 考了变体：

问，看了这个图（一模一样）怎么办？答案记不得了

Question 36 考了变体：（我的记性真的不错哦☺）

What option must be added to the program to obtain a data set containing Spearman statistics?

答案：outs=estimates

这个很好记：Pearson—outp=…

Spearman---outs=….

Question 37,38 没有考

Question 39 考了变体：SAS 又玩我们了

A non-contributing predictor variable (Pr > |t| =0.658) is eliminated from an existing multiple linear regression model.
What will be the result?

答案是 a decrease in R-square

Quesiton 40 又考了类似变体

也是给了一个公式，然后说其中一个 predictor 是 missing value，问我们预测值多少： 答案是：无法预测，因为 missing value

Question 41；

这个基本一样，但是把选择题换成了填空题，大家参考上面 R-squared 的公式自己算吧

答案 A

Question 42，完全原题

Question 43.考了变体：

一模一样的图表，但是问这个 dataset 一共多少样本 答案是：99+1=100

Question 44.

完全原题

Question 45：

完全原题

Question 46

完全原题

Question 47：

考了变体，Based on the SBC statistic, which model is the champion model?

到时候自己找 SBC 最小的那个就 ok 啦~SAS 歧视我们智商啊。。。

Question 48

一模一样原题，就是 7

Question 49

一模一样原题

Question 50

一模一样原题

Question 51：

一模一样原题，就是增加了一句话：salary 是 1000 为单位

选 B

Question 52:

一模一样原题，C 肯定错，如果是 C，没有删除 response value 的情况下 rerun 则得到一个新模型

Question 53

一模一样原题

Question 54

一模一样原题

Question 55

一模一样原题

Question 56

一模一样原题

Question 57

考了变体，SAS 又 2 了…

Which of the following describes a discordant pair of observations in the LOGISTIC procedure?

答案是 An observation with the event has a lower predicted probability than the observation without the event.

Question 58

没有考

Question 59

考了原题

Question 60

一模一样原题

Question 61

这个考了变体，问的是下列几个 statistic 中，哪一个是越大越好，自然选了 Adjusted R-squared

Question 62

没有考

Question 63

没有考

Question 64

没有考，不过这题很有意思，

首先，statistically significant， 我们排除了 Home

然后比较 Estimate 那项的绝对值

绝对值越大，越重要，选 C

Question 65

考了原题

哦哦，还有一题，考得是类似

proc surveyselect data=frame out=sample sampsize=800 outall;

run;

的语句，重点：sampsize=800 没有括号，别写成 sampsize=（800）！！！