《迴歸分析》

試題評析	今年高考迴歸分析以分析結果之解讀爲主，計算部分不難，雖然如此，但因爲解釋分析結果之意義需要清楚地把想法寫在考卷上，若敘述不清楚仍然會被扣分。最後第三大題有點像在考名詞解釋，又在測驗考生是否能夠把自己的想法用文字清楚地呈現出來。本卷基本分數爲70分。

一、抽樣某公司業務員之週業績（sales）與其每週工作時數（work）和年資（experience, 以年計），做兩個迴歸分析如下：

分析一：

Dependent Variable：sales

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	1	95.23381	95.23381	0.91	0.3473
Error	31	3241.31165	104.55844		
Corrected Total	32	3336.54545			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	144.77530	8.52019	16.99	<.0001
work	1	0.17402	0.18234	0.95	0.3473

分析二

Dependent Variable：sales

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr>F
Model	2	2186.59454	1093.29727	28.52	<.0001
Error	30	1149.95092	38.33170		
Corrected Total	32	3336.54545			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>|t|
Intercept	1	104.05180	7.55046	13.78	<.0001
Work	1	0.43469	0.11590	3.75	0.0008
experience	1	1.57153	0.21276	7.39	<.0001

(一)就分析一，當每週工作時數為45小時，估計週業績之平均數及變異數。（10分）

(二)比較分析一與分析二，每週工作時數與業績之關係有何差異？你會採用那個模式？為什麼？（$\alpha=0.05$）（15分）

(三)就分析二，在固定年資水準下，如工作時數增加一小時，檢定平均業績之增加量是否低於

0.5？列出H_0及H_1，並檢定之。（$\alpha=0.025$）（10分）

（四）就分析二，如年資的單位由年改為月，那些數值會改變？並求改變後之值。（10分）

考點命中	(一)高點出版《迴歸分析熱門題庫》，趙治勳老師編著，頁2-15~2-16。 (二)高點出版《迴歸分析熱門題庫》，趙治勳老師編著，頁2-42~2-43。 (三)高點出版《迴歸分析熱門題庫》，趙治勳老師編著，頁2-42~2-43。 (四)高點出版《迴歸分析熱門題庫》，趙治勳老師編著，頁2-26~2-27。

答：

設Y表週業績，X_1表每週工作時數，X_2表年資

假設迴歸模型：$Y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+\varepsilon_i$，$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

(一)樣本迴歸線：$\hat{y}=0.17402x_1$用以估計母體迴歸線：$E(Y)=\beta_1x_1$

故當每週工作時數$x_1=45$時，週業績之平均數估計值為$\hat{y}=0.17402(45)=7.8309$，而週業績之變異數$(V(Y)=V(\varepsilon)=\sigma^2)$估計值為$MSE=104.55844$。

(註：題目沒有給截距項之估計值,未知模型中是否有考慮截距項)

(二)在分析一中，有關每週工作時數X_1之迴歸係數T檢定之p-value=0.3473，在顯著水準$\alpha=0.05$下，我們沒有足夠證據去推論每週工作時數X_1對週業績Y之線性關係是存在的。

在分析二中，有關每週工作時數X_1之迴歸係數邊際T檢定之p-value=0.0008，在顯著水準$\alpha=0.05$下，我們有足夠證據去推論當模型考慮了年資X_2後，每週工作時數X_1對週業績Y之線性關係是存在的。

若以T檢定結論評估，分析二之模型較分析一佳，但是在分析二中之邊際T檢定是模型已經考慮了年資X_2下之檢定，若兩個自變數X_1，X_2間存在高度線性關係的話，此檢定是無法真實反映出每週工作時數X_1對週業績Y之影響能力是否顯著的，建議利用偏相關係數去評估。

(三)

$H_0:\beta_1\geq 0.5$ vs $H_1:\beta_1<0.5$

T.S.: $T=\dfrac{\hat{\beta}_1-0.5}{S(\hat{\beta}_1)}\sim t_{(30)}$

R.R.: Reject H_0 at $\alpha=0.025$ if $T^*<-t_{(30)0.025}=-\sqrt{F_{0.05(1,30)}}=2.042$

$\because T^*=\dfrac{0.43469-0.5}{0.11590}=-0.5635$ $\quad\therefore don'trejectH_0$

我們沒有足夠證據去推論平均業績之增量低於0.5。

(四)令原單位下之自變數為X_2，改變單位後為X_2^*，即$X_2^*=12X_2$

$\hat{\beta}_2^*=\dfrac{1}{12}\hat{\beta}_2=\dfrac{1}{12}(1.57153)=0.131$

$S(\hat{\beta}_2^*)=(\dfrac{1}{12})^2S(\hat{\beta}_2)=(\dfrac{1}{12})^2(0.21276)=0.0014775$ ($\because MSE$沒有變)

二、自四所高中隨機抽樣學生參加英文檢定考試，令Y＝成績，X1＝在校英文成績，X2＝每週讀英文之時數，（X3, X4, X5）＝（1, 0, 0）為甲校學生，（X3, X4, X5）＝（0, 1, 0）為乙校學生，（X3, X4, X5）＝（0, 0, 1）為丙校學生，（X3, X4, X5）＝（0, 0, 0）為丁校學生，得以下迴歸分析資訊。

模式：$Y=\beta_0+\beta_1X1+\beta_2X2+\beta_3X3+\beta_4X4+\beta_5X5+\varepsilon$

Analysis of Variance

Source	DF	Sum of Squares	Meam Square	F Value	Pr>F
Model	5	6368.47	1273.69	72.99	<.0001
X1		2768.78			
X2|X1		1.88			
X3|X1,X2		18.35			
X4|X1,X2,X3		3126.22			
X5|X1,X2,X3,X4		453.24			
Error	74	1291.33	17.45		
Corrected Total	79	7659.80			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr>| t |
Interecpt	1	42.680	6.530	6.54	<.0001
X1	1	0.504	0.071	7.10	<.0001
X2	1	0.059	0.310	0.19	0.8509
X3	1	-0.610	1.323	-0.46	0.6459
X4	1	-12.291	1.338	-9.19	<.0001
X5	1	6.827	1.340	5.10	<.0001

(一)就上述資訊：

1. 寫出對Y影響最大及次大之解釋變數，並請述明理由。（5分）
2. 某一甲校學生，在校英文成績為75，每週讀英文之時間為5小時，請預測其英文檢定成績。（5分）
3. 檢定 $H_0:\beta_3=\beta_4=\beta_5=0$ vs. H_1：非H_0。（$\alpha=0.05$）（12分）

(二)經由模式選取過程，得估計之模型為 $\hat{Y}=43.36+0.49X1-11.97X4+7.16X5$ 及MSR=2121.42。

1. 求此模型之修正後判定係數（adjusted coefficient of determination）。（10分）
2. 請解釋X5的係數7.16之涵義。（8分）

考點命中	(一)高點出版《迴歸分析熱門題庫》，趙治勳老師編著，頁2-42~2-43。 (二)高點出版《迴歸分析熱門題庫》，趙治勳老師編著，頁2-42及2-55~ 2-56。

答：

假設 $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

(一)1.以邊際T檢定作爲準則下，代表學校(甲乙丙丁)之虛擬變數X4之|t值|最大(9.19) ，故在其他自變數(在校英文成績X1,每週讀英文時數X2)已考慮在模型之下，自變數學校對應變數Y之影響力最大，而在校英文成績X1之|t值|爲次大(7.10) ，故在其他自變數(在校英文成績X1,學校)已考慮在模型之下，自變數在校英文成績X1對應變數Y之影響力次大。

(註：邊際T檢定是已經在其他自變數已經考慮在模型下之檢定，若自變數間存在高度線性關係的話，此檢定是無法真實反映出自變數對應變數之影響能力是否顯著，建議利用偏相關係數去評估自變數對應變數之影響力。)

2. $\hat{y}=42.68+0.504(75)+0.059(5)-0.61(1)-12.291(0)+6.827(0)=80.165$

故其英文檢定成績預測值爲80.165分。

3. $H_0:\beta_3=\beta_4=\beta_5=0$ vs $H_1: notH_0$

T.S.： $F=\dfrac{SSR(X_3X_4X_5\mid X_1X_2)/3}{SSE(X_1X_2X_3X_4X_5)/74}\sim F_{(3,74)}$

(註：分母自由度題目沒有給，我們是從考卷上所提供之查表值猜測是74)

R.R.： Reject H_0 at $\alpha=0.05$ if $F^*>F_{(3,74)0.05}=2.72$

$$\because F^*=\frac{SSR(X_3X_4X_5\mid X_1X_2)/3}{SSE(X_1X_2X_3X_4X_5)/74}$$

$$=\frac{[SSR(X_3\mid X_1X_2)+SSR(X_4\mid X_1X_2X_3)+SSR(X_5\mid X_1X_2X_3X_4)]/3}{MSE}$$

$$=\frac{[18.35+3126.22+453.24]/3}{1273.69/72.99}=68.725$$

$$\therefore rejectH_0$$

我們有足夠證據去推論至少一個 $\beta_i\neq 0, i=3,4,5$。

(二)

1.

ANOVA TABLE				
source	SS	d.f.	MS	F
Reg	6364.26	3	2121.42	$F^*=124.453$
Error	1295.51	76	17.046	
Total	7659.77	79		

故 $R^2_{adj}=1-\dfrac{1295.51/76}{7659.77/79}=0.824$

2.丙校學生： $\hat{y}=43.36+0.49x1-11.97(0)+7.16(1)=50.52+0.49x1$

乙校學生： $\hat{y}=43.36+0.49x1-11.97(1)+7.16(0)=31.39+0.49x1$

在固定英文成績X1下，丙校學生比乙校學生在英文檢定之平均成績上多19.13分。

三、迴歸分析中：

(一)如應變數Y不服從常態分配，該如何處置？（5分）

(二)如槓桿值（leverage, h）過大時，表示為何？（5分）

(三)修正後判定係數（adjusted coefficient of determination）有何用處？（5分）

考點命中	
	(一)高點出版《迴歸分析熱門題庫》，趙治勳老師編著，頁2-53。
	(二)高點出版《迴歸分析熱門題庫》，趙治勳老師編著，頁2-47。
	(三)高點出版《迴歸分析熱門題庫》，趙治勳老師編著，頁2-42。

答：

(一)1.在研究成本可以允許之情況下，增加樣本數至大樣本，就可以使用中央極限定理逼近常態分配，如此有關迴歸係數之顯著性檢定及信賴區間仍然可以獲得逼近之推論結果。

2.經由數據轉換使之近似常態分配，常用之轉換方法有Box-Cox transformation。

(二)槓桿值爲矩陣 $X_i^T(X^TX)X_i$ 上對角線之值，通常用 h_{ii} 表示，若第i個樣本之槓桿值 h_{ii} 大，表示該樣本之數據偏離其他樣本之數據，研究者應該再進一步討論原因，且研究該樣本影響迴歸分析之程度，以免導致分析結論之偏差。

(三)當自變數個數增加時，複判定係數 R^2 必然會增加或相等。其實，若研究者希望得到一個 R^2 解釋能力足夠

之迴歸模型時，只要增加自變數即可，但加入之變數是否真正擁有顯著之解釋能力就不得而知。

有鑑於此，統計學家發展出調整後之複判定係數 R^2_{adj} ，其想法是利用自由度去調整 R^2 之增量。定義 $R^2_{adj}=1-\frac{SSE/n-k-1}{SST/n-1}$ ，可得 $R^2_{adj}=1-(1-R^2)\frac{n-1}{n-k-1}$ ，如此，當自變數個數增加時，$(1-R^2)\downarrow$ 但 $\frac{n-1}{n-k-1}\uparrow$ ，一降一升就可達到平衡的作用。