المملكة العربية السعودية
جامعة الملك سعود
كلية العلوم
قسم الإحصاء وبحوث العمليات

مشكلة الارتباطات الخطية المشتركة المتعددة
Multicollinearity Problem

عمل الطالب:
سعود بن ناصر المانع
الرقم الجامعي: 438103392

إشراف:
أ.د. عبدالله بن عبدالكريم الشيحة

الفصل الدراسي الثاني 1443 هـ

مشروع تخرج مقدم لاستكمال متطلبات درجة البكالوريوس في قسم الإحصاء وبحوث العمليات

الفهرس

المقدمة

بسم الله الرحمن الرحيم والصلاة والسلام على أشرف الأنبياء والمرسلين نبينا محمد عليه أفضل الصلاة وأتم التسليم أما بعد:
تحليل الانحدار هو أداة إحصائية لاستكشاف العلاقة بين متغيرين كميين او أكثر للتنبؤ بأحد المتغيرات استنادا الى قيم المتغير او المتغيرات الأخرى، وفي تحليل الانحدار المتعدد يهتم المرء غالبا بطبيعة وأهمية العلاقات بين المتغيرات المستقلة والمتغير التابع، وبما أن موضوعنا عن مشكلة الارتباطات الخطية المشتركة فيجب علينا قبل ذلك شرح تحليل الانحدار المتعدد

أهداف البحث

- تعريف مؤشرات مشكلة الارتباطية الخطية المشتركة المتعددة.
- التعرف على مشاكل الارتباطية الخطية المشتركة المتعددة وإيجاد الحلول لها.
- تنفيذ بعض التطبيقات على مشاكل الارتباطية الخطية المشتركة المتعددة.

الفصل الأول
الانحدار الخطي المتعدد
Multiple regression

تحليل الانحدار الخطي المتعدد هو الأداة الإحصائية المستخدمة على أوسع نطاق من بين كافة الأدوات الإحصائية.
ويمكن الاستفادة من نماذج الانحدار الخطي المتعدد التي سنصفها في هذا البحث في تحليل البيانات.

نموذج الانحدار الخطي المتعدد
(Linear multiple regression model)

لنفرض أن المتغير التابع هو y والمتغيرات المستقلة هي $x_1, x_2, \dots, x_k$. ان نموذج الانحدار الخطي المتعدد يأخذ الصيغة التالية:

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_k X_{ik} + e_i$$
$$i = 1,2,..,n \quad , \quad k = 1,2,..,n$$

المعلمة β_0 هي قاطع المستوى. ونسمي أحيانًا β_j بمعامل الانحدار الجزئي لأنه يقيس التغيير المتوقع في Y لكل وحدة تغيير في x_j عندما تظل قيم المتغيرات المستقلة الأخرى ثابتة.
تسمى المعالم $\beta_j = 1,2,3, \dots, k$ بمعاملات الانحدار. يصف هذا النموذج المستوى الفائق في فضاء البعد k لمتغيرات الانحدار $\{x_j\}$.
غالبًا ما تُستخدم نماذج الانحدار الخطي المتعددة كدوال تقريبية لوصف العلاقة الدالية بين متغير الاستجابة والمتغيرات التابعة.
يمكن أيضًا تحليل النماذج التي تشتمل على تأثيرات التداخلات البينية من خلال نماذج الانحدار الخطي المتعدد.
بشكل عام، يمكن تعريف التداخل البيني على انه التأثير الناتج عن أحد المتغيرات المستقلة عند تغيير مستويات المتغيرات المستقلة الأخرى.

الفصل الثاني
مشكلة الارتباطية الخطية المشتركة المتعددة
Multicollinearity

في مسائل الانحدار الخطي المتعدد، نتوقع وجود علاقة بين متغير الاستجابة Y والمتغيرات المستقلة $x_j = 1,2,3, \dots, k$. وفي معظم مسائل الانحدار المتعدد، نجد أن هناك أيضًا علاقات بين المتغيرات المستقلة. وفي حالة كون هذه العلاقات قوية، فإن مشكلة الارتباطية الخطية المشتركة المتعددة تكون موجودة.
ويمكن أن يكون لهذه المشكلة تأثيرات خطيرة على تقديرات معاملات الانحدار مما يترتب عليها عدم دقة في الاستدلالات الإحصائية المبنية على النموذج المقدر.
ويمكن توضيح تأثيرات مشكلة الارتباطية الخطية المشتركة المتعددة من خلال المشاكل الرئيسية التي تبرز بصورة تقليدية عندما تكون المتغيرات المستقلة المعتبرة في نموذج الانحدار مرتبطة فيما بينها ارتباطا عاليا.
وفيما يأتي سوف نستعرض بعض من هذه التأثيرات:

1-اضافة او حذف متغير مستقل يترتب عليه تغير شديد في مقدرات معاملات الانحدار.

2-يتغير مجموع المربعات الإضافي المرافق لكل متغير مستقل اعتمادا على المتغيرات المستقلة الأخرى المشمولة في النموذج.

3-تصبح الانحرافات المعيارية لمقدرات معاملات الانحدار كبيرة عندما تكون المتغيرات المستقلة في نموذج الانحدار مرتبطة فيما بينها ارتباط عاليا. مما يؤدي إلى تغير كبير في قيمة مقدرات المعالم بالتغيرات الطفيفة على البيانات. (عدم استقرار مقدرات المعالم)

4-قد لا تكون جميع معاملات الانحدار المقدرة فرديا مهمة إحصائيا بالرغم من أن هناك علاقة ذات دلاله إحصائية بين المتغير التابع والمتغيرات المستقلة مجتمعة.

ويمكن توضيح تأثيرات الارتباطية الخطية المشتركة المتعددة بسهولة.
لنفرض ان المصفوفة X هي مصفوفة التصميم ولنفرض ان المصفوفة C هي المصفوفة القطرية للعناصر القطرية للمصفوفة $(X'X)^{-1}$ ويمكن كتابة العناصر القطرية للمصفوفة $C = (X'X)^{-1}$:

$$C_j \frac{1}{\left(1-R_j^2\right)} \; j = 1,2, \dots, k$$

حيث R_j^2 هو معامل التحديد المتعدد الناتج عن انحدار x_j على متغيرات الانحدار الأخرى. يمكننا التفكير في R_j^2 كمقياس للعلاقة بين x_j وعوامل الانحدار الأخرى.
ومن الواضح أنه كلما زادت قوة الارتباطية الخطية لـ x_j على المتغيرات المستقلة الاخرى، كلما زادت قيمة R_j^2 .

عامل تضخم التباين (Variance inflation factor)

إحدى الطرق الرسمية المستخدمة على نطاق واسع للكشف عن وجود مشكلة الارتباطية الخطية المشتركة المتعددة هي استخدام عوامل تضخم التباين (VIF) وتقيس هذه العوامل مدى تضخم تباينات معاملات الانحدار المقدرة بالمقارنة مع حالة عدم وجود صلة خطية بين المتغيرات المستقلة.

$$\text{VIF } (\beta_j) = \frac{1}{(1-R_j^2)} \; j = 1,2\dots,k$$

هذه العوامل هي مقاييس مهمة لمعرفة مدى وجود ارتباطية خطية مشتركة متعددة. إذا كانت أعمدة مصفوفة التصميم Xمتعامدة، فإن عوامل الانحدار تكون غير مرتبطة تمامًا، وستكون عوامل تضخم التباين كلها تساوي الواحد $(R_j^2 = 1)$.
لذلك، يشير أي عامل تضخم تباين (VIF) يتجاوز 1 إلى مستوى معين من الارتباطات الخطية المشتركة المتعددة في البيانات.
على الرغم من أن تقديرات معاملات الانحدار غير دقيقة للغاية عند وجود الارتباطات الخطية المشتركة المتعددة، فقد تظل معادلة النموذج المفروض مفيدة للتنبؤ بقيم المتغير التابع.
على سبيل المثال، افترض أننا نرغب في توقع مشاهدات جديدة على الاستجابة، إذا كانت هذه التنبؤات عبارة عن استقراء في المنطقة الأصلية من الفضاء x حيث تكون الارتباطية الخطية المشتركة المتعددة سارية المفعول، فسيتم الحصول على

تنبؤات مرضية غالبًا في حين قد يكون تقدير المعلمة β_j ضعيفًا، وبالتالي يمكن تقدير الدالة $\sum_{j=1}^{k} \beta_i \, x_{ij}$ جيدًا.
من ناحية أخرى، إذا كان التنبؤ بملاحظات جديدة يتطلب استقراءً يتجاوز المنطقة الأصلية من الفضاء x حيث تم جمع البيانات، فإننا نتوقع عمومًا الحصول على نتائج سيئة.
تنشأ الارتباطية الخطية المشتركة المتعددة لعدة أسباب. وقد تحدث هذه المشكلة (على سبيل المثال) عندما يقوم المحلل بجمع البيانات بحيث يثبت القيد الخطي تقريبًا بين أعمدة المصفوفة.X فلو كان لدينا أربع متغيرات مستقلة أحدهما تركيب خطي في المتغيرات الأخرى، فسيظل هذا القيد موجودًا دائمًا، وعادة قد لا يعرف المحلل أن هذا القيد موجودا.

طرق الكشف عن وجود مشكلة الارتباطية الخطية المشتركة المتعددة:
يمكن الكشف عن وجود ارتباطية خطية مشتركة متعددة بعدة طرق:

1-تعتبر عوامل تضخم التباين مقاييس مفيدة جدًا للارتباطية الخطية المشتركة المتعددة. كلما زاد عامل تضخم التباين، زادت شدة الارتباطية الخطية المشتركة المتعددة.
اقترح بعض المؤلفين أنه إذا تجاوز أي عامل تضخم تباين القيمة 10، تعد الارتباطية الخطية المشتركة المتعددة مشكلة.

2. إذا كان اختبار F لأهمية الانحدار مهمًا، ولكن الاختبارات على كل عامل من معاملات الانحدار الفردية ليست مهمة، فقد يكون هناك مشكلة ارتباطية خطية مشتركة متعددة.

علاج مشكلة الارتباطية الخطية المشتركة المتعددة:
تم اقتراح العديد من التدابير العلاجية لحل مشكلة الارتباطية الخطية المشتركة المتعددة.
غالبًا ما يُقترح زيادة البيانات بمشاهدات جديدة مصممة خصيصًا لتفكيك الارتباطية الخطية التقريبية الموجودة حاليًا. ومع ذلك، فقد يكون هذا مستحيلًا في بعض الأحيان لأسباب اقتصادية أو بسبب القيود المادية. وكخيار آخر، نقوم بحذف أحد او بعض المتغيرات المستقلة من النموذج، ولكن هذا النهج له عيب في تجاهل المعلومات الواردة في المتغيرات المحذوفة.
نظرًا لأن الارتباطية الخطية المشتركة المتعددة تؤثر بشكل أساسي على استقرار معاملات الانحدار، فقد يكون من المفيد تقدير هذه المعالم بطريقة ما أقل حساسية للارتباطات الخطية المشتركة المتعددة من المربعات الصغرى العادية.

الفصل الثالث
تطبيقات على مشكلة الارتباطية الخطية المشتركة المتعددة
Applications on Multicollinearity

تطبيق 1

سوف نستخدم برنامج R لإجراء تحليل الانحدار الخطي المتعدد لدراسة وجود مشكلة الارتباطية الخطية المشتركة المتعددة للبيانات التالية:

Observation Number	x_1	x_2	x_3	x_4	x_5	y
1	3	3	3	3	0	0.787
2	8	30	8	8	0	0.293
3	3	6	6	6	0	1.710
4	4	4	4	12	0	0.203
5	8	7	6	5	0	0.806
6	10	20	5	5	0	4.713
7	8	6	3	3	25	0.607
8	6	24	4	4	25	9.107
9	4	10	12	4	25	9.210
10	16	12	8	4	25	1.365
11	3	10	8	8	25	4.554
12	8	3	3	3	25	0.293
13	3	6	3	3	50	2.252
14	3	8	8	3	50	9.167
15	4	8	4	8	50	0.694
16	5	2	2	2	50	0.379
17	2	2	2	3	50	0.485
18	10	15	3	3	50	3.345
19	15	6	2	3	50	0.208
20	15	6	2	3	75	0.201
21	10	4	3	3	75	0.329
22	3	8	2	2	75	4.966
23	6	6	6	4	75	1.362
24	2	3	8	6	75	1.515
25	3	3	8	8	75	0.751

George Applied Statistics and Probability for Engineers, Sixth Edition, by Douglas C. Montgomery and C. Runger (table E12-15) page 535

البيانات عبارة عن مقارنة مقدار الفولت y لدى محولين كهربائيين وخمس معاملات للانحدار:

Y: مقدار الفولت (متغير الاستجابة)

X1: طول جهاز NMOS

X2: عرض جهازNMOS

X3: طول جهازPMOS

X4: عرض جهازPMOS

X5: درجة الحرارة

والان نقوم بعمل مصفوفة الترابط للبيانات السابقة:

```
> cor(grad_data3)
         x1          x2          x3          x4          x5          y
x1  1.000000000  0.2478995 -0.1994436 -0.2588852 -0.002109001 -0.25079140
x2  0.247899530  1.0000000  0.2662058  0.1598969 -0.387679001  0.40246702
x3 -0.199443591  0.2662058  1.0000000  0.3992473 -0.190361799  0.42738103
x4 -0.258885196  0.1598969  0.3992473  1.0000000 -0.336189478 -0.15848683
x5 -0.002109001 -0.3876790 -0.1903618 -0.3361895  1.000000000 -0.05771861
y  -0.250791397  0.4024670  0.4273810 -0.1584868 -0.057718614  1.00000000
```

نلاحظ أن قيم معاملات الارتباط الخطي الثنائي بين كل متغيرين من المتغيرات المستقلة صغيرة وهذه إشارة جيدة لعدم وجود مشكلة الارتباطية الخطية المشتركة المتعددة، حيث انه كلما كانت العلاقة الخطية بين المتغيرات المستقلة قوية كلما ارتفعت فرصة وجود مشكلة ارتباطية خطية مشتركة متعددة، ولكن لن نقول بوجود او عدم وجود مشكلة ارتباطية خطية مشتركة متعددة بين المتغيرات المستقلة الا بعد اجراء اختبار عامل تضخم التباين.

والان سنقوم بتحليل الانحدار الخطي المتعدد للبيانات وأيضا اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> model4 = lm(y~x1 + x2 + x3 + x4 + x5, data = grad_data3)
> summary(model4)
> library(car)
> vif(model4)
```

```
Coefficients:
             Estimate     Std. Error   t value    Pr(>|t|)
(Intercept)  2.85473    1.86922   1.527   0.1432
x1           -0.29047    0.11742  -2.474   0.0230 *
x2            0.20572    0.07506   2.741   0.0130 *
x3            0.45444    0.18768   2.421   0.0256 *
x4           -0.59419    0.21253  -2.796   0.0115 *
x5            0.00464    0.01817   0.255   0.8012
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.196 on 19 degrees of freedom
Multiple R-squared:  0.5584,      Adjusted R-squared:  0.4422
F-statistic: 4.805 on 5 and 19 DF,  p-value: 0.005239
```

	X1	X2	X3	X4	X5
VIF	1.2249	1.3795	1.2995	1.3658	1.2951

نلاحظ انه بعد القيام باختبار عامل تضخم التباين تبين انه لا يوجد مشكلة ارتباطية خطية مشتركة متعددة بين البيانات وبالتالي هذه البيانات لا تحتوي على مشكلة ارتباطيه خطية مشتركة متعددة.

تطبيق 2

سوف نستخدم برنامج R لإجراء تحليل الانحدار الخطي المتعدد لدراسة وجود مشكلة الارتباطية الخطية المشتركة المتعددة للبيانات التالية:

Observation Number	y	x_1	x_2	x_3	x_4	x_5	x_6
1	4540	2140	20640	30250	205	1732	99
2	4315	2016	20280	30010	195	1697	100
3	4095	1905	19860	29780	184	1662	97
4	3650	1675	18980	29330	164	1598	97
5	3200	1474	18100	28960	144	1541	97
6	4833	2239	20740	30083	216	1709	87
7	4617	2120	20305	29831	206	1669	87
8	4340	1990	19961	29604	196	1640	87
9	3820	1702	18916	29088	171	1572	85
10	3368	1487	18012	28675	149	1522	85
11	4445	2107	20520	30120	195	1740	101
12	4188	1973	20130	29920	190	1711	100
13	3981	1864	19780	29720	180	1682	100
14	3622	1674	19020	29370	161	1630	100
15	3125	1440	18030	28940	139	1572	101
16	4560	2165	20680	30160	208	1704	98
17	4340	2048	20340	29960	199	1679	96
18	4115	1916	19860	29710	187	1642	94
19	3630	1658	18950	29250	164	1576	94
20	3210	1489	18700	28890	145	1528	94
21	4330	2062	20500	30190	193	1748	101
22	4119	1929	20050	29960	183	1713	100
23	3891	1815	19680	29770	173	1684	100
24	3467	1595	18890	29360	153	1624	99
25	3045	1400	17870	28960	134	1569	100
26	4411	2047	20540	30160	193	1746	99
27	4203	1935	20160	29940	184	1714	99
28	3968	1807	19750	29760	173	1679	99
29	3531	1591	18890	29350	153	1621	99
30	3074	1388	17870	28910	133	1561	99
31	4350	2071	20460	30180	198	1729	102
32	4128	1944	20010	29940	186	1692	101
33	3940	1831	19640	29750	178	1667	101
34	3480	1612	18710	29360	156	1609	101
35	3064	1410	17780	28900	136	1552	101
36	4402	2066	20520	30170	197	1758	100
37	4180	1954	20150	29950	188	1729	99
38	3973	1835	19750	29740	178	1690	99
39	3530	1616	18850	29320	156	1616	99
40	3080	1407	17910	28910	137	1569	100

Applied Statistics and Probability for Engineers, Sixth Edition, by Douglas C. Montgomery and George C. Runger (table E12-14) page 534

البيانات عبارة عن
Y: محرك دفع تيربو نفاث
X1: سرعة الدوران الأساسية
X2: سرعة الدوران الثانوية
X3: معدل تدفق الوقود
X4: الضغط
X5: درجة حرارة العادم
X6: درجة الحرارة المحيطة وقت الاختبار

والان نقوم بعمل مصفوفة الترابط للبيانات السابقة:

```
> cor(grad_data2)
         y          x1          x2          x3          x4          x5          x6
y  1.0000000  0.99500990 0.97604801 0.9287925  0.9951122 0.8716925 -0.14744067
x1 0.9950099  1.00000000 0.98517759 0.9503955  0.9939651 0.8940304 -0.07274465
x2 0.9760480  0.98517759 1.00000000 0.9723041  0.9718586 0.9267470  0.01713334
x3 0.9287925  0.95039552 0.97230414 1.0000000  0.9203139 0.9757503  0.21646594
x4 0.9951122  0.99396510 0.97185865 0.9203139  1.0000000 0.8519012 -0.15342505
x5 0.8716925  0.89403037 0.92674701 0.9757503  0.8519012 1.0000000  0.30176412
x6 -0.1474407 -0.07274465 0.01713334 0.2164659 -0.1534251 0.3017641  1.00000000
```

نلاحظ أن معظم قيم معاملات الارتباطات الخطية بين المتغيرات المستقلة عالية مما يشير الى وجود مشكلة الارتباطية الخطية المشتركة المتعددة، على سبيل المثال فإن معامل الارتباط الخطي بين x4 وx1 يساوي 0.9939 وهذه القيمة عالية مما يثير قلقنا من احتمالية وجود مشكلة ارتباطية خطية مشتركة متعددة وسوف نقوم بإجراء اختبار تضخم التباين لمعرفة وجود المشكلة ام لا.

الخطوة الأولى:

والان سنقوم بتحليل الانحدار الخطي المتعدد للبيانات وأيضا اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> model2 = lm(y~ x1 + x2 + x3 + x4 + x5 + x6, data = grad_data2)
> summary(model2)
> library(car)
> vif(model2)
```

```
Coefficients:
             Estimate      Std. Error   t value   Pr(>|t|)
(Intercept) -4.738e+03  2.445e+03  -1.938   0.061213 .
x1           1.119e+00  2.865e-01   3.904   0.000441 ***
x2          -3.018e-02  3.823e-02   -0.789   0.435478
x3           2.306e-01  1.180e-01    1.954   0.059231 .
x4           3.850e+00  2.686e+00    1.433   0.161246
x5           8.219e-01  3.508e-01    2.343   0.025298 *
x6          -1.695e+01  2.620e+00   -6.468   2.45e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.51 on 33 degrees of freedom
Multiple R-squared:  0.9977,      Adjusted R-squared:  0.9972
F-statistic:  2350 on 6 and 33 DF,  p-value: < 2.2e-16
```

	X1	X2	X3	X4	X5	X6
VIF	289.1131	71.8295	168.0510	219.9722	32.4090	8.4773

نظرا لان معظم قيم عوامل تضخم التباين (VIF > 10) مما يشير الى وجود مشكلة الارتباطية الخطية المشتركة المتعددة.

نلاحظ انه يوجد مشكلة ارتباطية خطية مشتركة متعددة لدى عدة متغيرات بقيم عالية جدا لذا سوف نقوم بإجراء اختبار الأهمية للمتغير x2 لأن قيمة ال p-value له هي الاعلى بمقدار 0.435 ولأن هذه القيمة أكبر من مستوى الدلالة 0.05 ولذلك سوف نحذف المتغير x2 ثمن نقوم بإعادة تحليل الانحدار الخطي المتعدد مره أخرى.

الخطوة الثانية:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات (x1,x3,x4,x5,x6) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
         y       x1       x3       x4       x5        x6
y  1.0000000 0.99500990 0.9287925 0.9951122 0.8716925 -0.14744067
x1 0.9950099 1.00000000 0.9503955 0.9939651 0.8940304 -0.07274465
x3 0.9287925 0.95039552 1.0000000 0.9203139 0.9757503  0.21646594
x4 0.9951122 0.99396510 0.9203139 1.0000000 0.8519012 -0.15342505
x5 0.8716925 0.89403037 0.9757503 0.8519012 1.0000000  0.30176412
x6 -0.1474407 -0.07274465 0.2164659 -0.1534251 0.3017641  1.00000000
```

```
> model2 = lm(y~ x1  + x3 + x4 + x5 + x6, data = grad_data)
> summary(model2)
> library(car)
> vif(model2)
```

```
Coefficients:
             Estimate       Std. Error   t value   Pr(>|t|)
(Intercept) -3982.1058  2236.9356  -1.780    0.083989 .
x1              1.0964       0.2835       3.867    0.000473 ***
x3              0.1843       0.1018       1.810    0.079193 .
x4              3.7456       2.6681       1.404    0.169432
x5              0.8343       0.3484       2.394    0.022308 *
x6            -16.2781      2.4658      -6.602    1.44e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.36 on 34 degrees of freedom
Multiple R-squared:  0.9976,         Adjusted R-squared:  0.9973
F-statistic:  2852 on 5 and 34 DF,  p-value: < 2.2e-16
```

	X1	X3	X4	X5	X6
VIF	286.3559	126.4891	219.4439	32.3435	7.5923

بما ان نسبة R-squared لا تزال مرتفعة فيمكننا حذف المتغير x2.
نلاحظ انه نسب مصفوفة الارتباط بين المتغيرات لاتزال مرتفعة جدا بين جميع المتغيرات تقريبا وأيضا نتائج اختبار عامل تضخم التباين لاتزال مرتفعة ولذلك سوف نقوم بحذف المتغير x4 لأنه يمتلك اعلى قيمه p-value ونقوم بإجراء اختبار أهميته:

الخطوة الثالثة:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات (x1,x3,x5,x6) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
        y         x1        x3        x5         x6
y  1.0000000 0.99500990 0.9287925 0.8716925 -0.14744067
x1 0.9950099 1.00000000 0.9503955 0.8940304 -0.07274465
x3 0.9287925 0.95039552 1.0000000 0.9757503  0.21646594
x5 0.8716925 0.89403037 0.9757503 1.0000000  0.30176412
x6 -0.1474407 -0.07274465 0.2164659 0.3017641  1.00000000
```

```
> model2 = lm(y~ x1  + x3 + x5 + x6, data = grad_data)
> summary(model2)
> library(car)
> vif(model2)
```

```
Coefficients:
             Estimate      Std. Error     t value     Pr(>|t|)
(Intercept) -4280.1645  2257.5088  -1.896   0.0662 .
x1              1.4420     0.1426       10.114   6.30e-12 ***
x3              0.2098     0.1016       2.066    0.0463 *
x5              0.6467     0.3262       1.982    0.0553 .
x6            -17.5103     2.3360       -7.496   8.86e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.72 on 35 degrees of freedom
Multiple R-squared:  0.9975,      Adjusted R-squared:  0.9972
F-statistic:  3468 on 4 and 35 DF,  p-value: < 2.2e-16
```

	X1	X3	X5	X6
VIF	70.4701	122.4522	27.5876	6.6303

بما ان نسبة R-squared لا تزال مرتفعة فيمكننا حذف المتغير x4.
نلاحظ انه نسب مصفوفة الارتباط بين المتغيرات لاتزال مرتفعة جدا بين جميع المتغيرات تقريبا وأيضا نتائج اختبار عامل تضخم التباين لاتزال مرتفعة، ولكن انخفضت بشكل ملحوظ ومع بقاء نسبة R-squared كما هي تقريبا ولذلك سوف نقوم بحذف المتغير x5 لأنه يمتلك اعلى قيمه p-value ونقوم بإجراء اختبار أهميته:

الخطوة الرابعة:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات (x1,x3,x6) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
          y          x1          x3          x6
y  1.0000000 0.99500990 0.9287925 -0.14744067
x1 0.9950099 1.00000000 0.9503955 -0.07274465
x3 0.9287925 0.95039552 1.0000000  0.21646594
x6 -0.1474407 -0.07274465 0.2164659  1.00000000
```

```
> model2 = lm(y~ x1  + x3 + x6, data = grad_data)
> summary(model2)
> library(car)
> vif(model2)
```

```
Coefficients:
             Estimate    Std. Error    t value    Pr(>|t|)
(Intercept) -7018.4846  1856.8069  -3.780  0.00057 ***
x1              1.3631      0.1424       9.574     1.96e-11 ***
x3              0.3443      0.0786       4.380      9.80e-05 ***
x6            -17.8438     2.4229      -7.365       1.10e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.79 on 36 degrees of freedom
Multiple R-squared:  0.9972,      Adjusted R-squared:  0.997
F-statistic:  4275 on 3 and 36 DF,  p-value: < 2.2e-16
```

	X1	X3	X6
VIF	64.8919	67.5959	6.5959

بما ان نسبة R-squared لا تزال مرتفعة فيمكننا حذف المتغير x5.
نلاحظ انه نسب مصفوفة الارتباط بين المتغيرات لاتزال مرتفعة بين جميع المتغيرات، ولكن تقريبا المتغير x6 يملك اقل نسبة ترابط بين المتغيرات الاخرى وأيضا نتائج اختبار عامل تضخم التباين لاتزال مرتفعة، ولكن انخفضت بشكل ملحوظ ومع بقاء نسبة R-squared كما هي تقريبا ولذلك سوف نقوم إجراء جميع الاختبارات الممكنة لكي نحصل على اعلى نسبة من R-squared وأيضا عدم وجود مشكلة الارتباطية الخطية المشتركة المتعددة لذا سوف نقوم بحذف أولا المتغير x6 ونقوم بإجراء اختبار أهميته:

الخطوة الخامسة:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات (x1,x3) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
        y         x1        x3
y  1.0000000 0.9950099 0.9287925
x1 0.9950099 1.0000000 0.9503955
x3 0.9287925 0.9503955 1.0000000
```

```
> model2 = lm(y~ x1 + x3, data = grad_data)
> summary(model2)
> library(car)
> vif(model2)
```

```
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept) 5281.16157 1267.28255   4.167   0.000178 ***
x1             2.32471     0.08868     26.215   < 2e-16 ***
x3            -0.18864     0.04792    -3.936    0.000352 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.4 on 37 degrees of freedom
Multiple R-squared:  0.993,      Adjusted R-squared:  0.9926
F-statistic:  2618 on 2 and 37 DF,  p-value: < 2.2e-16
```

	X1	X3
VIF	10.3360	10.3360

نلاحظ ان لاتزال نسب مصفوفة الترابط بين المتغيرات مرتفعة وأيضا لا تزال مشكلة الارتباطية الخطية المشتركة موجودة لذا سوف نحاول متغير اخر وهو x1 مع إعادة المتغير x6 للاختبار.

الخطوة السادسة:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات (x3,x6) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
          y         x3         x6
y  1.0000000 0.9287925 -0.1474407
x3 0.9287925 1.0000000  0.2164659
x6 -0.1474407 0.2164659  1.0000000
```

```
> model2 = lm(y~ x3 +x6, data = grad_data)
> summary(model2)
> library(car)
> vif(model2)
```

```
Coefficients:
             Estimate     Std. Error    t value   Pr(>|t|)
(Intercept) -2.459e+04  5.282e+02  -46.54   <2e-16 ***
x3           1.091e+00  1.816e-02   60.08   <2e-16 ***
x6          -3.912e+01  1.795e+00  -21.79   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.62 on 37 degrees of freedom
Multiple R-squared:  0.9901,      Adjusted R-squared:  0.9895
F-statistic:  1845 on 2 and 37 DF,  p-value: < 2.2e-16
```

	X3	X6
VIF	1.049161	1.049161

نلاحظ نتائج اختبار عامل تضخم التباين اعطى نتائج تدل على عدم وجود مشكلة ارتباطية خطية مشتركة متعددة، ولكن يوجد اختبار أخير بخذف المتغير x3 ممكن يعطيني نتيجة أفضل ل R-squared لذا سوف نقوم اختبار اهميته.

الخطوة السابعة:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات (x1,x6) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
          y        x1        x6
y  1.0000000 0.99500990 -0.14744067
x1 0.9950099 1.00000000 -0.07274465
x6 -0.1474407 -0.07274465 1.00000000
```

```
> model2 = lm(y~ x1 +x6, data = grad_data)
> summary(model2)
> library(car)
> vif(model2)
```

```
Coefficients:
             Estimate   Std. Error   t value   Pr(>|t|)
(Intercept) 1103.27736  121.94030   9.048   6.52e-11 ***
x1          1.98197     0.02163     91.637   < 2e-16 ***
x6          -8.07277    1.15524     -6.988   2.94e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.94 on 37 degrees of freedom
Multiple R-squared:  0.9957,     Adjusted R-squared:  0.9955
F-statistic:  4292 on 2 and 37 DF,  p-value: < 2.2e-16
```

	X1	X6
VIF	1.00532	1.00532

	R^2
X1, X6	0.9957
X3, X6	0.9901

نظرا لان قيم جميع معاملات التضخم (VIF <10)
تبين عدم وجود مشكلة ارتباطية خطية مشتركة متعددة وهذا الاختبار لديه نتيجة متساوية تقريبا ل R-squared الذي يحوي على جميع المتغيرات المستقلة والنموذج النهائي لذا سوف نعتمد x1,x6.

النموذج	R^2
النموذج الاصلي	0.9977
النموذج النهائي	0.9957

وكما نلاحظ ان معامل الارتباط الخطي بين x1,x6 يساوي 0.070- وهذه القيمة ضئيلة جدا مما يعني عدم وجود مشكلة ارتباطية خطية مشتركة متعددة بين x1,x6.

تطبيق 3

سوف نستخدم برنامج R لإجراء تحليل الانحدار الخطي المتعدد لدراسة وجود مشكلة الارتباطية الخطية المشتركة المتعددة للبيانات التالية:

Observation Number	y	x_1	x_2	x_3
1	0.22200	7.3	0.0	0.0
2	0.39500	8.7	0.0	0.3
3	0.42200	8.8	0.7	1.0
4	0.43700	8.1	4.0	0.2
5	0.42800	9.0	0.5	1.0
6	0.46700	8.7	1.5	2.8
7	0.44400	9.3	2.1	1.0
8	0.37800	7.6	5.1	3.4
9	0.49400	10.0	0.0	0.3
10	0.45600	8.4	3.7	4.1
11	0.45200	9.3	3.6	2.0
12	0.11200	7.7	2.8	7.1
13	0.43200	9.8	4.2	2.0
14	0.10100	7.3	2.5	6.8
15	0.23200	8.5	2.0	6.6
16	0.30600	9.5	2.5	5.0
17	0.09230	7.4	2.8	7.8
18	0.11600	7.8	2.8	7.7
19	0.07640	7.7	3.0	8.0
20	0.43900	10.3	1.7	4.2
21	0.09440	7.8	3.3	8.5
22	0.11700	7.1	3.9	6.6
23	0.07260	7.7	4.3	9.5
24	0.04120	7.4	6.0	10.9
25	0.25100	7.3	2.0	5.2
26	0.00002	7.6	7.8	20.7

George Applied Statistics and Probability for Engineers, Sixth Edition, by Douglas C. Montgomery and C. Runger (table E12-13) page 531

البيانات عبارة عن

Y: محلول الكسر الجزئي

X1: ذوبان التشتت الجزئي

X2: ذوبان ثنائي القطب الجزئي

X3: ذوبان روابط الهيدروجين الجزيئية

كل هذه المتغيرات مع درجة حرارة ثابتة

والان نقوم بعمل مصفوفة الترابط بين المتغيرات:

```
cor(grad_data1) <
          y          x1          x2          x3
y  1.0000000  0.7620911 -0.4670590 -0.8117565
x1 0.7620911  1.0000000 -0.3717451 -0.4925686
x2 -0.4670590 -0.3717451  1.0000000  0.7243160
x3 -0.8117565 -0.4925686  0.7243160  1.0000000
```

نلاحظ أن قيم معاملات الارتباط الخطي الثنائي بين كل متغيرين من المتغيرات المستقلة صغيرة وهذه إشارة جيدة لعدم وجود مشكلة الارتباطية الخطية المشتركة المتعددة، حيث انه كلما كانت العلاقة الخطية بين المتغيرات المستقلة قوية كلما ارتفعت فرصة وجود مشكلة ارتباطية خطية مشتركة متعددة، ولكن لن نقول بوجود او عدم وجود مشكلة ارتباطية خطية مشتركة متعددة بين المتغيرات المستقلة الا بعد اجراء اختبار عامل تضخم التباين.

والان سنقوم بتحليل الانحدار الخطي المتعدد للبيانات وأيضا اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> model1 = lm(y~ x1 + x2 + x3 , data = grad_data1)
> summary(model1)
> library(car)
> vif(model1)
```

```
Coefficients:
             Estimate      Std. Error    t value  Pr(>|t|)
(Intercept) -0.369306    0.143646  -2.571   0.0174 *
x1           0.086510    0.016003   5.406   1.98e-05 ***
x2           0.024412    0.010226   2.387   0.0260 *
x3          -0.028577    0.004477  -6.383   2.01e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06581 on 22 degrees of freedom
Multiple R-squared:  0.8667,      Adjusted R-squared:  0.8486
F-statistic: 47.69 on 3 and 22 DF,  p-value: 8.535e-10
```

	X1	X1	X3
VIF	1.321170	2.104951	2.395189

نلاحظ انه بعد القيام باختبار عامل تضخم التباين تبين انه لا يوجد مشكلة ارتباطية خطية مشتركة متعددة بين البيانات وبالتالي هذه البيانات لا تحتوي على مشكلة ارتباطيه خطية مشتركة متعددة.

تطبيق **4**

سوف نستخدم برنامج R لإجراء تحليل الانحدار الخطي المتعدد لدراسة وجود مشكلة الارتباطية الخطية المشتركة المتعددة للبيانات التالية:

Team	W	AVG	R	H	2B	3B	HR
Chicago	99	0.262	741	1450	253	23	200
Boston	95	0.281	910	1579	339	21	199
LA Angels	95	0.27	761	1520	278	30	147
New York	95	0.276	886	1552	259	16	229
Cleveland	93	0.271	790	1522	337	30	207
Oakland	88	0.262	772	1476	310	20	155
Minnesota	83	0.259	688	1441	269	32	134
Toronto	80	0.265	775	1480	307	39	136
Texas	79	0.267	865	1528	311	29	260
Baltimore	74	0.269	729	1492	296	27	189
Detroit	71	0.272	723	1521	283	45	168
Seattle	69	0.256	699	1408	289	34	130
Tampa Bay	67	0.274	750	1519	289	40	157
Kansas City	56	0.263	701	1445	289	34	126
St. Louis	100	0.27	805	1494	287	26	170
Atlanta	90	0.265	769	1453	308	37	184
Houston	89	0.256	693	1400	281	32	161
Philadelphia	88	0.27	807	1494	282	35	167
Florida	83	0.272	717	1499	306	32	128
New York	83	0.258	722	1421	279	32	175
San Diego	82	0.257	684	1416	269	39	130
Milwaukee	81	0.259	726	1413	327	19	175
Washington	81	0.252	639	1367	311	32	117
Chicago	79	0.27	703	1506	323	23	194
Arizona	77	0.256	696	1419	291	27	191
San Francisco	75	0.261	649	1427	299	26	128
Cincinnati	73	0.261	820	1453	335	15	222
Los Angeles	71	0.253	685	1374	284	21	149
Colorado	67	0.267	740	1477	280	34	150
Pittsburgh	67	0.259	680	1445	292	38	139

يتبع...

بسبب كثرة المتغيرات في البيانات قسمناها على جدولين مع ثبات أسماء الفرق والمتغير التابع (W)

Team	W	RBI	BB	SO	SB	GIDP	LOB	OBP
Chicago	99	713	435	1002	137	122	1032	0.322
Boston	95	863	653	1044	45	135	1249	0.357
LA Angels	95	726	447	848	161	125	1086	0.325
New York	95	847	637	989	84	125	1264	0.355
Cleveland	93	760	503	1093	62	128	1148	0.334
Oakland	88	739	537	819	31	148	1170	0.33
Minnesota	83	644	485	978	102	155	1109	0.323
Toronto	80	735	486	955	72	126	1118	0.331
Texas	79	834	495	1112	67	123	1104	0.329
Baltimore	74	700	447	902	83	145	1103	0.327
Detroit	71	678	384	1038	66	137	1077	0.321
Seattle	69	657	466	986	102	115	1076	0.317
Tampa Bay	67	717	412	990	151	133	1065	0.329
Kansas City	56	653	424	1008	53	139	1062	0.32
St. Louis	100	757	534	947	83	127	1152	0.339
Atlanta	90	733	534	1084	92	146	1114	0.333
Houston	89	654	481	1037	115	116	1136	0.322
Philadelphia	88	760	639	1083	116	107	1251	0.348
Florida	83	678	512	918	96	144	1181	0.339
New York	83	683	486	1075	153	103	1122	0.322
San Diego	82	655	600	977	99	122	1220	0.333
Milwaukee	81	689	531	1162	79	137	1120	0.331
Washington	81	615	491	1090	45	130	1137	0.322
Chicago	79	674	419	920	65	131	1133	0.324
Arizona	77	670	606	1094	67	132	1247	0.332
San Francisco	75	617	431	901	71	147	1093	0.319
Cincinnati	73	784	611	1303	72	116	1176	0.339
Los Angeles	71	653	541	1094	58	139	1135	0.326
Colorado	67	704	509	1103	65	125	1197	0.333
Pittsburgh	67	656	471	1092	73	130	1193	0.322

George Applied Statistics and Probability for Engineers, Sixth Edition, by Douglas C. Montgomery and C. Runger (table E12-16) page 536

البيانات عبارة عن احصائيات لدوري كرة القاعدة الأمريكي للمحترفين عام 2005

الانتصارات W Wins

معدل ضربات الكرة AVG Batting average

الركض الى احد الجولات R Runs

الضربات H Hits

جولتين 2B Doubles

ثلاث جولات 3B Triples

ضربة ساحقة HR Home runs

ضربات مساعدة لتحقيق جولة RBI Runs batted in
عدد اللاعبين المحققين لجولة BB Walks
خروج اللاعب من القاعدة SO Strikeouts
الجولات الغير محسوبة SB
ضربة أرضية تؤدي للخروج من الجولة GIDP
عدد اللاعبين المتبقين في الجولة بعد انتهاء اللعب LOB Left on base
معدل الوصول الى الجولة OBP On-base percentage

والان نقوم بعمل مصفوفة الترابط للبيانات السابقة:

```
> cor(grad_data)
             W          AVG           R           H          B2          B3
W    1.00000000  0.29331820  0.46728089  0.306485236 -0.03525997 -0.33465780
AVG  0.29331820  1.00000000  0.69994633  0.962140537  0.17518219 -0.02379701
R    0.46728089  0.69994633  1.00000000  0.764282082  0.24731601 -0.35140316
H    0.30648524  0.96214054  0.76428208  1.000000000  0.18496559 -0.05202669
B2  -0.03525997  0.17518219  0.24731601  0.184965586  1.00000000 -0.27099384
B3  -0.33465780 -0.02379701 -0.35140316 -0.052026692 -0.27099384  1.00000000
HR   0.34251742  0.39533781  0.71263036  0.496221055  0.24497242 -0.48623857
RBI  0.47306417  0.68356890  0.99462423  0.758880701  0.26520566 -0.37531919
BB   0.37489812  0.07559290  0.51950626  0.087179247  0.14823412 -0.41793955
SO  -0.21404138 -0.24860727  0.14403415 -0.241276895  0.28610288 -0.09694214
SB   0.23322157  0.04007596 -0.05859257  0.001907258 -0.53312844  0.26081983
GIDP -0.14159998  0.08240286 -0.22198040  0.044510498  0.17607615 -0.06528689
LOB  0.23362069  0.17155504  0.35620318  0.184262600  0.10983350 -0.24152783
OBP  0.46265964  0.60991676  0.78271891  0.575580200  0.21179377 -0.35649592
             HR          RBI          BB          SO          SB          GIDP
W    0.34251742  0.4730642  0.37489812 -0.21404138  0.233221569 -0.14159998
AVG  0.39533781  0.6835689  0.07559290 -0.24860727  0.040075964  0.08240286
R    0.71263036  0.9946242  0.51950626  0.14403415 -0.058592574 -0.22198040
H    0.49622105  0.7588807  0.08717925 -0.24127689  0.001907258  0.04451050
B2   0.24497242  0.2652057  0.14823412  0.28610288 -0.533128438  0.17607615
B3  -0.48623857 -0.3753192 -0.41793955 -0.09694214  0.260819828 -0.06528689
HR   1.00000000  0.7480195  0.30886421  0.36864807 -0.053785581 -0.23830756
RBI  0.74801950  1.0000000  0.52175760  0.15485242 -0.069710902 -0.22174807
BB   0.30886421  0.5217576  1.00000000  0.35839844 -0.211061972 -0.21998562
SO   0.36864807  0.1548524  0.35839844  1.00000000 -0.147754022 -0.36479305
SB  -0.05378558 -0.0697109 -0.21106197 -0.14775402  1.000000000 -0.42256945
GIDP -0.23830756 -0.2217481 -0.21998562 -0.36479305 -0.422569454  1.00000000
LOB  0.17893837  0.3667929  0.85572984  0.21212018 -0.298361646 -0.17957567
OBP  0.43413528  0.7751499  0.81882762  0.16209546 -0.147485223 -0.11526848
            LOB       OBP
W    0.2336207  0.4626596
AVG  0.1715550  0.6099168
R    0.3562032  0.7827189
H    0.1842626  0.5755802
B2   0.1098335  0.2117938
B3  -0.2415278 -0.3564959
HR   0.1789384  0.4341353
RBI  0.3667929  0.7751499
BB   0.8557298  0.8188276
SO   0.2121202  0.1620955
SB  -0.2983616 -0.1474852
GIDP -0.1795757 -0.1152685
LOB  1.0000000  0.7563281
OBP  0.7563281  1.0000000
```

نلاحظ أن بعض قيم معاملات الارتباطات الخطية بين المتغيرات المستقلة عالية مما يشير الى وجود مشكلة الارتباطية الخطية المشتركة المتعددة، على سبيل المثال فإن معامل الارتباط الخطي بين H وAVG يساوي 0.9621 وهذه القيمة عالية مما يثير قلقنا من احتمالية وجود مشكلة ارتباطية خطية مشتركة متعددة وسوف نقوم بإجراء اختبار تضخم التباين لمعرفة وجود المشكلة ام لا.

الخطوة الأولى:

والان سنقوم بتحليل الانحدار الخطي المتعدد للبيانات وأيضا اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> model5 = lm(W~AVG  + R+ H+ B2 + B3+ HR +RBI +BB + SO +SB+GIDP+ LOB +OBP,data =
grad_data)
> summary(model5)
> library(car)
> vif(model5)
```

```
Coefficients:
             Estimate     Std. Error    t value  Pr(>|t|)
(Intercept)  2.586e+01  1.335e+02   0.194   0.8488
AVG          -1.468e+03  2.384e+03  -0.616   0.5467
R            8.273e-02  3.688e-01   0.224   0.8253
H            3.605e-03  2.953e-01   0.012   0.9904
B2           9.198e-02  1.166e-01   0.789   0.4417
B3           1.093e-01  3.753e-01   0.291   0.7747
HR           1.615e-01  1.268e-01   1.274   0.2210
RBI          -8.213e-02  3.982e-01  -0.206   0.8392
BB           -6.716e-02  1.655e-01  -0.406   0.6902
SO           -7.543e-02  3.162e-02  -2.386   0.0297 *
SB           1.075e-01  9.609e-02   1.119   0.2796
GIDP         -3.448e-02  2.562e-01  -0.135   0.8946
LOB          -3.056e-02  1.258e-01  -0.243   0.8112
OBP          1.571e+03  1.611e+03   0.975   0.3440
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 9.655 on 16 degrees of freedom
Multiple R-squared:  0.5619,      Adjusted R-squared:  0.2059
F-statistic: 1.578 on 13 and 16 DF,  p-value: 0.1919
```

	AVG	R	H	B2	B3	HR	RBI
VIF	93.7011	186.2889	75.1891	2.0852	2.4796	6.1274	202.6133
	BB	SO	SB	GIDP	LOB	OBP	
VIF	44.48	3.1027	3.0891	3.075	18.6088	80.6357	

نظرا لان معظم قيم عوامل تضخم التباين (VIF > 10) مما يشير الى وجود مشكلة الارتباطية الخطية المشتركة المتعددة.
نلاحظ انه يوجد مشكلة ارتباطية خطية مشتركة متعددة لدى عدة متغيرات بقيم عالية جدا لذا سوف نقوم بإجراء اختبار الأهمية للمتغير H لأن قيمة ال p-value له هي الاعلى بمقدار 0.9904 ولأن هذه القيمة أكبر من مستوى الدلالة 0.05 ولذلك سوف نحذف المتغير H ثم نقوم بإعادة تحليل الانحدار الخطي المتعدد مره أخرى.

الخطوة الثانية:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات
(AVG,R,B2,B3,HR,RBI,BB,SO,SB,GIDP,LOB,OBP) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
              W         AVG           R          B2          B3          HR
W     1.00000000  0.29331820  0.46728089 -0.03525997 -0.33465780  0.34251742
AVG   0.29331820  1.00000000  0.69994633  0.17518219 -0.02379701  0.39533781
R     0.46728089  0.69994633  1.00000000  0.24731601 -0.35140316  0.71263036
B2   -0.03525997  0.17518219  0.24731601  1.00000000 -0.27099384  0.24497242
B3   -0.33465780 -0.02379701 -0.35140316 -0.27099384  1.00000000 -0.48623857
HR    0.34251742  0.39533781  0.71263036  0.24497242 -0.48623857  1.00000000
RBI   0.47306417  0.68356890  0.99462423  0.26520566 -0.37531919  0.74801950
BB    0.37489812  0.07559290  0.51950626  0.14823412 -0.41793955  0.30886421
SO   -0.21404138 -0.24860727  0.14403415  0.28610288 -0.09694214  0.36864807
SB    0.23322157  0.04007596 -0.05859257 -0.53312844  0.26081983 -0.05378558
GIDP -0.14159998  0.08240286 -0.22198040  0.17607615 -0.06528689 -0.23830756
LOB   0.23362069  0.17155504  0.35620318  0.10983350 -0.24152783  0.17893837
OBP   0.46265964  0.60991676  0.78271891  0.21179377 -0.35649592  0.43413528
             RBI          BB          SO          SB        GIDP         LOB
W     0.4730642  0.3748981 -0.21404138  0.23322157 -0.14159998  0.2336207
AVG   0.6835689  0.0755929 -0.24860727  0.04007596  0.08240286  0.1715550
R     0.9946242  0.5195063  0.14403415 -0.05859257 -0.22198040  0.3562032
B2    0.2652057  0.1482341  0.28610288 -0.53312844  0.17607615  0.1098335
B3   -0.3753192 -0.4179396 -0.09694214  0.26081983 -0.06528689 -0.2415278
HR    0.7480195  0.3088642  0.36864807 -0.05378558 -0.23830756  0.1789384
RBI   1.0000000  0.5217576  0.15485242 -0.06971090 -0.22174807  0.3667929
BB    0.5217576  1.0000000  0.35839844 -0.21106197 -0.21998562  0.8557298
SO    0.1548524  0.3583984  1.00000000 -0.14775402 -0.36479305  0.2121202
SB   -0.0697109 -0.2110620 -0.14775402  1.00000000 -0.42256945 -0.2983616
GIDP -0.2217481 -0.2199856 -0.36479305 -0.42256945  1.00000000 -0.1795757
LOB   0.3667929  0.8557298  0.21212018 -0.29836165 -0.17957567  1.0000000
OBP   0.7751499  0.8188276  0.16209546 -0.14748522 -0.11526848  0.7563281
          OBP
W     0.4626596
AVG   0.6099168
R     0.7827189
B2    0.2117938
B3   -0.3564959
HR    0.4341353
RBI   0.7751499
BB    0.8188276
SO    0.1620955
SB   -0.1474852
GIDP -0.1152685
LOB   0.7563281
OBP   1.0000000
```

```
> model5 = lm(W~AVG      + R+ B2 + B3+ HR +RBI +BB +SO +SB+GIDP+ LOB +OBP,data = grad_data)
> summary(model5)
> library(car)
> vif(model5)
```

```
Coefficients:
             Estimate    Std. Error    t value    Pr(>|t|)
(Intercept)  2.638e+01  1.228e+02   0.215   0.8324
AVG          -1.443e+03  1.256e+03  -1.149   0.2663
R            8.256e-02  3.575e-01   0.231   0.8201
B2           9.195e-02  1.131e-01   0.813   0.4274
B3           1.098e-01  3.617e-01   0.304   0.7652
HR           1.614e-01  1.228e-01   1.314   0.2062
RBI          -8.042e-02  3.617e-01  -0.222   0.8267
BB           -6.701e-02  1.601e-01  -0.419   0.6808
SO           -7.549e-02  3.033e-02  -2.489   0.0235 *
SB           1.077e-01  9.219e-02   1.168   0.2588
GIDP         -3.330e-02  2.304e-01  -0.145   0.8868
LOB          -2.961e-02  9.584e-02  -0.309   0.7611
OBP          1.558e+03  1.221e+03   1.276   0.2191
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.366 on 17 degrees of freedom
Multiple R-squared:  0.5619,          Adjusted R-squared:  0.2526
F-statistic: 1.817 on 12 and 17 DF,  p-value: 0.1267
```

	AVG	R	B2	B3	HR	RBI
VIF	27.6292	186.0360	2.0845	2.4470	6.1089	177.5940
	BB	SO	SB	GIDP	LOB	OBP
VIF	44.2258	3.0346	3.0216	2.6426	11.4717	49.2581

بما ان نسبة R-squared لا تزال مرتفعة فيمكننا حذف المتغير H.
نلاحظ انه نسب مصفوفة الارتباط بين المتغيرات انخفضت، ولكن لا تزال مرتفعة بين بعض المتغيرات وأيضا نتائج اختبار عامل تضخم التباين لاتزال مرتفعة ولذلك سوف نقوم بحذف المتغير GIDP لأنه يمتلك اعلى قيمه p-value ونقوم بإجراء اختبار أهميته:

الخطوة الثالثة:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات
(AVG,R,B2,B3,HR,RBI,BB,SO,SB,GIDP,LOB,OBP) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
           W          AVG          R           B2          B3          HR
W   1.00000000 0.29331820 0.46728089 -0.03525997 -0.33465780 0.34251742
AVG 0.29331820 1.00000000 0.69994633 0.17518219 -0.02379701 0.39533781
R   0.46728089 0.69994633 1.00000000 0.24731601 -0.35140316 0.71263036
B2  -0.03525997 0.17518219 0.24731601 1.00000000 -0.27099384 0.24497242
B3  -0.33465780 -0.02379701 -0.35140316 -0.27099384 1.00000000 -0.48623857
HR  0.34251742 0.39533781 0.71263036 0.24497242 -0.48623857 1.00000000
RBI 0.47306417 0.68356890 0.99462423 0.26520566 -0.37531919 0.74801950
BB  0.37489812 0.07559290 0.51950626 0.14823412 -0.41793955 0.30886421
SO  -0.21404138 -0.24860727 0.14403415 0.28610288 -0.09694214 0.36864807
SB  0.23322157 0.04007596 -0.05859257 -0.53312844 0.26081983 -0.05378558
LOB 0.23362069 0.17155504 0.35620318 0.10983350 -0.24152783 0.17893837
OBP 0.46265964 0.60991676 0.78271891 0.21179377 -0.35649592 0.43413528
          RBI         BB          SO          SB          LOB         OBP
W   0.4730642 0.3748981 -0.21404138 0.23322157 0.2336207 0.4626596
AVG 0.6835689 0.0755929 -0.24860727 0.04007596 0.1715550 0.6099168
R   0.9946242 0.5195063 0.14403415 -0.05859257 0.3562032 0.7827189
B2  0.2652057 0.1482341 0.28610288 -0.53312844 0.1098335 0.2117938
B3  -0.3753192 -0.4179396 -0.09694214 0.26081983 -0.2415278 -0.3564959
HR  0.7480195 0.3088642 0.36864807 -0.05378558 0.1789384 0.4341353
RBI 1.0000000 0.5217576 0.15485242 -0.06971090 0.3667929 0.7751499
BB  0.5217576 1.0000000 0.35839844 -0.21106197 0.8557298 0.8188276
SO  0.1548524 0.3583984 1.00000000 -0.14775402 0.2121202 0.1620955
SB  -0.0697109 -0.2110620 -0.14775402 1.00000000 -0.2983616 -0.1474852
LOB 0.3667929 0.8557298 0.21212018 -0.29836165 1.0000000 0.7563281
OBP 0.7751499 0.8188276 0.16209546 -0.14748522 0.7563281 1.0000000
```

```
> model5 = lm(W~AVG      + R+ B2 + B3+ HR +RBI +BB +SO +SB+LOB +OBP,data = grad_data)
> summary(model5)
> library(car)
> vif(model5)
```

```
Coefficients:
```

```
                Estimate     Std. Error   t value Pr(>|t|)
(Intercept)  2.360e+01  1.179e+02   0.200   0.8436
AVG          -1.495e+03  1.170e+03  -1.278   0.2174
R             9.743e-02  3.330e-01   0.293   0.7731
B2            9.371e-02  1.093e-01   0.857   0.4027
B3            1.120e-01  3.514e-01   0.319   0.7535
HR            1.611e-01  1.194e-01   1.350   0.1939
RBI          -8.885e-02  3.471e-01  -0.256   0.8009
BB           -7.347e-02  1.495e-01  -0.491   0.6290
SO           -7.445e-02  2.866e-02  -2.598   0.0182 *
SB            1.161e-01  6.974e-02   1.665   0.1133
LOB          -2.221e-02  7.878e-02  -0.282   0.7812
OBP           1.557e+03  1.188e+03   1.311   0.2063
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.108 on 18 degrees of freedom
Multiple R-squared:  0.5613,        Adjusted R-squared:  0.2932
F-statistic: 2.094 on 11 and 18 DF,  p-value: 0.07913
```

	AVG	R	B2	B3	HR	RBI
VIF	25.3609	170.6295	2.0606	2.4425	6.1077	172.9708
	BB	SO	SB	LOB	OBP	
VIF	40.7791	2.8652	1.8282	8.1973	49.2558	

بما ان نسبة R-squared لا تزال مرتفعة فيمكننا حذف المتغير GIDP.
نلاحظ انه نسب مصفوفة الارتباط بين المتغيرات انخفضت، ولكن لا تزال مرتفعة بين بعض المتغيرات وأيضا نتائج اختبار عامل تضخم التباين لاتزال مرتفعة ولذلك سوف نقوم بحذف المتغير RBI لأنه يمتلك اعلى قيمه p-value ونقوم بإجراء اختبار أهميته:

الخطوة الرابعة:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات (AVG,R,B2,B3,HR,BB,SO,SB,LOB,OBP) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
           W          AVG           R          B2          B3          HR
W   1.00000000  0.29331820  0.46728089 -0.03525997 -0.33465780  0.34251742
AVG 0.29331820  1.00000000  0.69994633  0.17518219 -0.02379701  0.39533781
R   0.46728089  0.69994633  1.00000000  0.24731601 -0.35140316  0.71263036
B2  -0.03525997 0.17518219  0.24731601  1.00000000 -0.27099384  0.24497242
B3  -0.33465780 -0.02379701 -0.35140316 -0.27099384  1.00000000 -0.48623857
HR  0.34251742  0.39533781  0.71263036  0.24497242 -0.48623857  1.00000000
BB  0.37489812  0.07559290  0.51950626  0.14823412 -0.41793955  0.30886421
SO  -0.21404138 -0.24860727  0.14403415  0.28610288 -0.09694214  0.36864807
SB  0.23322157  0.04007596 -0.05859257 -0.53312844  0.26081983 -0.05378558
LOB 0.23362069  0.17155504  0.35620318  0.10983350 -0.24152783  0.17893837
OBP 0.46265964  0.60991676  0.78271891  0.21179377 -0.35649592  0.43413528
           BB          SO          SB         LOB         OBP
W   0.3748981 -0.21404138  0.23322157  0.2336207  0.4626596
AVG 0.0755929 -0.24860727  0.04007596  0.1715550  0.6099168
R   0.5195063  0.14403415 -0.05859257  0.3562032  0.7827189
B2  0.1482341  0.28610288 -0.53312844  0.1098335  0.2117938
B3  -0.4179396 -0.09694214  0.26081983 -0.2415278 -0.3564959
HR  0.3088642  0.36864807 -0.05378558  0.1789384  0.4341353
BB  1.0000000  0.35839844 -0.21106197  0.8557298  0.8188276
SO  0.3583984  1.00000000 -0.14775402  0.2121202  0.1620955
SB  -0.2110620 -0.14775402  1.00000000 -0.2983616 -0.1474852
LOB 0.8557298  0.21212018 -0.29836165  1.0000000  0.7563281
OBP 0.8188276  0.16209546 -0.14748522  0.7563281  1.0000000
```

```
> model5 = lm(W~AVG  + R+ B2 + B3+ HR +BB +SO +SB+LOB +OBP,data = grad_data)
> summary(model5)
> library(car)
> vif(model5)
```

```
Coefficients:
             Estimate   Std. Error   t value  Pr(>|t|)
(Intercept) 2.410e+01  1.149e+02   0.210   0.8362
AVG         -1.397e+03  1.078e+03  -1.296   0.2103
R            1.444e-02  7.399e-02   0.195   0.8473
B2           8.493e-02  1.012e-01   0.839   0.4120
B3           1.019e-01  3.404e-01   0.299   0.7680
HR           1.446e-01  9.785e-02   1.478   0.1559
BB          -6.246e-02  1.396e-01  -0.447   0.6596
SO          -7.199e-02  2.632e-02  -2.735   0.0132 *
SB           1.137e-01  6.737e-02   1.687   0.1079
LOB         -2.832e-02  7.320e-02  -0.387   0.7031
OBP          1.488e+03  1.128e+03   1.320   0.2027
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.881 on 19 degrees of freedom
Multiple R-squared:  0.5597,        Adjusted R-squared:  0.328
F-statistic: 2.416 on 10 and 19 DF,  p-value: 0.04706
```

	AVG	R	B2	B3	HR
VIF	22.6360	8.8633	1.8578	2.4115	4.3132
	BB	SO	SB	LOB	OBP
VIF	37.4042	2.5413	1.7945	7.4431	46.6991

بما ان نسبة R-squared لا تزال مرتفعة فيمكننا حذف المتغير RBI.

نلاحظ انه نسب مصفوفة الارتباط بين المتغيرات انخفضت، ولكن لا تزال مرتفعة بين بعض المتغيرات وأيضا نتائج اختبار عامل تضخم التباين لاتزال مرتفعة ولذلك سوف نقوم بحذف المتغير R لأنه يمتلك اعلى قيمه p-value ونقوم بإجراء اختبار أهميته:

الخطوة الخامسة:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات (AVG,B2,B3,HR,BB,SO,SB, LOB,OBP) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
           W         AVG          B2          B3          HR          BB
W   1.00000000  0.29331820 -0.03525997 -0.33465780  0.34251742  0.3748981
AVG 0.29331820  1.00000000  0.17518219 -0.02379701  0.39533781  0.0755929
B2  -0.03525997 0.17518219  1.00000000 -0.27099384  0.24497242  0.1482341
B3  -0.33465780 -0.02379701 -0.27099384  1.00000000 -0.48623857 -0.4179396
HR  0.34251742  0.39533781  0.24497242 -0.48623857  1.00000000  0.3088642
BB  0.37489812  0.07559290  0.14823412 -0.41793955  0.30886421  1.0000000
SO  -0.21404138 -0.24860727  0.28610288 -0.09694214  0.36864807  0.3583984
SB  0.23322157  0.04007596 -0.53312844  0.26081983 -0.05378558 -0.2110620
LOB 0.23362069  0.17155504  0.10983350 -0.24152783  0.17893837  0.8557298
OBP 0.46265964  0.60991676  0.21179377 -0.35649592  0.43413528  0.8188276
           SO          SB         LOB         OBP
W   -0.21404138  0.23322157  0.2336207  0.4626596
AVG -0.24860727  0.04007596  0.1715550  0.6099168
B2  0.28610288 -0.53312844  0.1098335  0.2117938
B3  -0.09694214  0.26081983 -0.2415278 -0.3564959
HR  0.36864807 -0.05378558  0.1789384  0.4341353
BB  0.35839844 -0.21106197  0.8557298  0.8188276
SO  1.00000000 -0.14775402  0.2121202  0.1620955
SB  -0.14775402  1.00000000 -0.2983616 -0.1474852
LOB 0.21212018 -0.29836165  1.0000000  0.7563281
OBP 0.16209546 -0.14748522  0.7563281  1.0000000
```

```
> model5 = lm(W~AVG+ B2 + B3+ HR +BB +SO +SB+LOB +OBP,data = grad_data)
> summary(model5)
> library(car)
> vif(model5)
```

```
Coefficients:
             Estimate   Std. Error   t value  Pr(>|t|)
(Intercept) 1.669e+01  1.059e+02   0.158   0.8763
AVG         -1.329e+03 9.942e+02  -1.336   0.1964
B2           8.410e-02  9.868e-02   0.852   0.4042
B3           1.130e-01  3.275e-01   0.345   0.7337
HR           1.550e-01  8.004e-02   1.937   0.0671 .
BB          -5.221e-02  1.262e-01  -0.414   0.6835
SO          -7.242e-02  2.559e-02  -2.830   0.0103 *
SB          1.114e-01  6.473e-02   1.721   0.1008
LOB         -3.652e-02  5.849e-02  -0.624   0.5395
OBP         1.497e+03  1.099e+03   1.362   0.1883
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.665 on 20 degrees of freedom
Multiple R-squared:  0.5588,        Adjusted R-squared:  0.3603
F-statistic: 2.815 on 9 and 20 DF,  p-value: 0.02583
```

	AVG	B2	B3	HR	BB
VIF	20.2403	1.8546	2.3444	3.0314	32.1097
	SO	SB	LOB	OBP	
VIF	2.5228	1.7403	4.9935	46.6170	

بما ان نسبة R-squared لا تزال مرتفعة فيمكننا حذف المتغير R.
نلاحظ انه نسب مصفوفة الارتباط بين المتغيرات انخفضت، ولكن لا تزال مرتفعة بين بعض المتغيرات وأيضا نتائج اختبار عامل تضخم التباين لاتزال مرتفعة ولذلك سوف نقوم بحذف المتغير B3 لأنه يمتلك اعلى قيمه p-value ونقوم بإجراء اختبار أهميته:

الخطوة السادسة:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات (AVG,B2,HR,BB,SO,SB, LOB,OBP) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
           W          AVG           B2          HR          BB          SO
W   1.00000000  0.29331820 -0.03525997  0.34251742  0.3748981 -0.2140414
AVG 0.29331820  1.00000000  0.17518219  0.39533781  0.0755929 -0.2486073
B2  -0.03525997 0.17518219  1.00000000  0.24497242  0.1482341  0.2861029
HR  0.34251742  0.39533781  0.24497242  1.00000000  0.3088642  0.3686481
BB  0.37489812  0.07559290  0.14823412  0.30886421  1.0000000  0.3583984
SO  -0.21404138 -0.24860727  0.28610288  0.36864807  0.3583984  1.0000000
SB  0.23322157  0.04007596 -0.53312844 -0.05378558 -0.2110620 -0.1477540
LOB 0.23362069  0.17155504  0.10983350  0.17893837  0.8557298  0.2121202
OBP 0.46265964  0.60991676  0.21179377  0.43413528  0.8188276  0.1620955
           SB          LOB          OBP
W   0.23322157  0.2336207  0.4626596
AVG 0.04007596  0.1715550  0.6099168
B2  -0.53312844 0.1098335  0.2117938
HR  -0.05378558 0.1789384  0.4341353
BB  -0.21106197 0.8557298  0.8188276
SO  -0.14775402 0.2121202  0.1620955
SB  1.00000000 -0.2983616 -0.1474852
LOB -0.29836165  1.0000000  0.7563281
OBP -0.14748522  0.7563281  1.0000000
```

```
> model5 = lm(W~AVG+ B2 + HR +BB +SO +SB+LOB +OBP,data = grad_data)
> summary(model5)
> library(car)
> vif(model5)
```

```
Coefficients:
             Estimate    Std. Error    t value   Pr(>|t|)
(Intercept)  1.744e+01  1.036e+02   0.168  0.86790
AVG          -1.178e+03  8.737e+02  -1.348  0.19202
B2           7.646e-02  9.413e-02   0.812  0.42573
HR           1.383e-01  6.241e-02   2.216  0.03787 *
BB           -4.444e-02  1.215e-01  -0.366  0.71827
SO           -6.798e-02  2.164e-02  -3.142  0.00492 **
SB           1.145e-01  6.272e-02   1.826  0.08209 .
LOB          -3.296e-02  5.636e-02  -0.585  0.56489
OBP          1.360e+03  1.003e+03   1.356  0.18962
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.481 on 21 degrees of freedom
Multiple R-squared:  0.5562,      Adjusted R-squared:  0.3872
F-statistic:  3.29 on 8 and 21 DF,  p-value: 0.01358
```

	AVG	B2	HR	BB	SO	SB	LOB	OBP
VIF	16.3129	1.7612	1.9240	31.0868	1.8827	1.7054	4.8380	40.5580

بما ان نسبة R-squared لا تزال مرتفعة فيمكننا حذف المتغير B3.
نلاحظ انه نسب مصفوفة الارتباط بين المتغيرات انخفضت، ولكن لا تزال مرتفعة بين بعض المتغيرات وأيضا نتائج اختبار عامل تضخم التباين لاتزال مرتفعة ولذلك سوف نقوم بحذف المتغير BB لأنه يمتلك اعلى قيمه p-value ونقوم بإجراء اختبار أهميته:

الخطوة السابعة:

الجدول التالي يحتوي على مصفوفة الترابط وايضا تحليل الانحدار الخطي المتعدد للمتغير التابع للمتغيرات (AVG,B2,HR,SO,SB, LOB,OBP) مع اختبار عامل تضخم التباين باستخدام برنامج ال R والجدول التالي يحوي أوامر ومخرجات برنامج ال R:

```
> cor(grad_data)
           W          AVG           B2          HR          SO          SB
W   1.00000000  0.29331820 -0.03525997  0.34251742 -0.2140414  0.23322157
AVG 0.29331820  1.00000000  0.17518219  0.39533781 -0.2486073  0.04007596
B2  -0.03525997 0.17518219  1.00000000  0.24497242  0.2861029 -0.53312844
HR  0.34251742  0.39533781  0.24497242  1.00000000  0.3686481 -0.05378558
SO  -0.21404138 -0.24860727 0.28610288  0.36864807  1.0000000 -0.14775402
SB  0.23322157  0.04007596 -0.53312844 -0.05378558 -0.1477540  1.00000000
LOB 0.23362069  0.17155504  0.10983350  0.17893837  0.2121202 -0.29836165
OBP 0.46265964  0.60991676  0.21179377  0.43413528  0.1620955 -0.14748522
          LOB       OBP
W   0.2336207  0.4626596
AVG 0.1715550  0.6099168
B2  0.1098335  0.2117938
HR  0.1789384  0.4341353
SO  0.2121202  0.1620955
SB  -0.2983616 -0.1474852
LOB 1.0000000  0.7563281
OBP 0.7563281  1.0000000
```

```
> model5 = lm(W~AVG+ B2 + HR+SO +SB+LOB +OBP,data = grad_data)
> summary(model5)
> library(car)
> vif(model5)
```

```
Coefficients:
             Estimate    Std. Error    t value Pr(>|t|)
(Intercept)  42.48529   76.17930   0.558   0.5827
AVG          -891.45250 380.15791  -2.345   0.0285 *
B2           0.07198     0.09147   0.787   0.4397
HR           0.13007     0.05704   2.280   0.0327 *
SO           -0.06648    0.02082  -3.193   0.0042 **
SB           0.11099     0.06074   1.827   0.0812 .
LOB          -0.04202    0.04960  -0.847   0.4060
OBP          1022.75362  385.49650   2.653   0.0145 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.313 on 22 degrees of freedom
Multiple R-squared:  0.5534,      Adjusted R-squared:  0.4113
F-statistic: 3.894 on 7 and 22 DF,  p-value: 0.006621
```

	AVG	B2	HR	SO	SB	LOB	OBG
VIF	6.2311	3.9017	1.6648	1.8147	1.6732	1.7314	3.2152

النموذج	R^2
النموذج الأصلي	0.5619
النموذج النهائي	0.5534

نظرا لان قيم جميع معاملات التضخم (VIF <10)
تبين عدم وجود مشكلة ارتباطية خطية مشتركة متعددة وهذا الاختبار لديه نتيجة متساوية تقريبا ل R-squared الذي يحوي على جميع المتغيرات المستقلة والنموذج النهائي لذا سوف نعتمد AVG,B2,HR,SO,SB, LOB,OBP.

وكما نلاحظ ان معامل الارتباط الخطي مثلا بين HR,SB يساوي 0.0537- وهذه القيمة ضئيلة جدا وأيضا اغلب معاملات الارتباط بين المتغيرات المستقلة الأخرى مما يعني عدم وجود مشكلة ارتباطية خطية مشتركة متعددة بين المتغيرات المستقلة.

```
> cor(grad_data)
           W         AVG          B2          HR          SO          SB
W   1.00000000  0.29331820 -0.03525997  0.34251742 -0.2140414  0.23322157
AVG 0.29331820  1.00000000  0.17518219  0.39533781 -0.2486073  0.04007596
B2  -0.03525997 0.17518219  1.00000000  0.24497242  0.2861029 -0.53312844
HR  0.34251742  0.39533781  0.24497242  1.00000000  0.3686481 -0.05378558
SO  -0.21404138 -0.24860727 0.28610288  0.36864807  1.0000000 -0.14775402
SB  0.23322157  0.04007596 -0.53312844 -0.05378558 -0.1477540  1.00000000
LOB 0.23362069  0.17155504  0.10983350  0.17893837  0.2121202 -0.29836165
OBP 0.46265964  0.60991676  0.21179377  0.43413528  0.1620955 -0.14748522
          LOB       OBP
W   0.2336207  0.4626596
AVG 0.1715550  0.6099168
B2  0.1098335  0.2117938
HR  0.1789384  0.4341353
SO  0.2121202  0.1620955
SB  -0.2983616 -0.1474852
LOB 1.0000000  0.7563281
OBP 0.7563281  1.0000000
```

الخلاصة

إذا كانت قيم معاملات الارتباط الخطي الثنائي بين كل متغيرين من المتغيرات المستقلة صغيرة، هذه إشارة جيدة لعدم وجود مشكلة الارتباطية الخطية المشتركة المتعددة، حيث انه كلما كانت العلاقة الخطية بين المتغيرات المستقلة قوية كلما ارتفعت فرصة وجود مشكلة ارتباطية خطية مشتركة متعددة، ولكن لن نقول بوجود او عدم وجود مشكلة ارتباطية خطية مشتركة متعددة بين المتغيرات المستقلة الا بعد اجراء اختبار عامل تضخم التباين.

اذا كانت احدى قيم عوامل تضخم التباين (VIF > 10) فإنه يشير الى وجود مشكلة الارتباطية الخطية المشتركة المتعددة.
إذا وجد مشكلة ارتباطية خطية مشتركة متعددة نقوم بإجراء اختبار الأهمية للمتغير الذي يملك أكبر قيمه من ال p-value وإذا كانت اكبر من قيمة الدلالة 0.05 فسوف نحذف المتغير المستقل ونعيد الاختبار مرة أخرى مع الاهتمام بقيمة R-squared.

من خلال كل ما سبق ذكره تبين ان اختبار عامل تضخم التباين (VIF) اختبار فعال للكشف عن مشكلة الارتباطية الخطية المشتركة المتعددة وهو يساعدنا على وجود او عدم وجود ترابط قوي بين المتغيرات المستقلة من خلال دراسة أهمية المتغير المستقل وقوة تأثيره على المتغيرات المستقلة الأخرى والمتغير التابع.

المراجع

١-جون نتر، وليام وازرمان، ميخائيل كتنر كتاب "نماذج إحصائية خطية تطبيقية: انحدار، تحليل تباين وتصاميم تجريبية" الجزء الأول (الانحدار)، مترجم للعربية بواسطة: أ.د. أنيس كنجو، أ.د. عبد الحميد الزيد، د. إبراهيم الواصل، د. الحسيني راضي. جامعة الملك سعود، كلية العلوم، قسم الإحصاء وبحوث العمليات. (2000م ، 1421هـ) النسخة الثالثة.

2-Applied Statistics and Probability for Engineers, Sixth Edition, by Douglas C. Montgomery and George C. Runger,2014.