رگرسیون در SAS – 4

در این درس به مسائلی فراتر از OLS می‌پردازیم. این درس تفاوت کمی از دیگر درس‌ها دارد ازاین‌جهت که مفاهیم متفاوتی که ممکن است برای شما جدید باشد را پوشش می‌دهد. این فرمت‌ها، فراتر از OLS، علاوه بر بسیاری از جستجوها برای OLS ابزار اضافه برای کار با مدل‌های خطی را برای شما تهیه می‌کند.

عناوین شامل روش‌های رگرسیون Robust، رگرسیون خطی اجباری، رگرسیون با داده‌های سانسور و بریده‌شده، رگرسیون خطای اندازه‌گیری و مدل‌های معادله چندگانه می‌باشد.

روش‌های رگرسیون Robust

به نظر می‌رسد به‌ندرت در مجموعه داده‌ها پیش‌فرض‌های رگرسیون برقرار باشد. می‌دانیم که برقرار نبودن پیش‌فرض‌ها باعث برآورد اریب ضرایب و مخصوصاً برآورد اریب خطای استاندارد می‌شود. به این دلیل به توسعه روش‌های رگرسیون Robust می‌پردازیم.

ایده پشت روش رگرسیونی Robust این است که در برآوردها تعدیلاتی را انجام دهیم که برخی معایب داده‌ها را در نظر بگیرند. قصد داریم سه روش Robust را بررسی کنیم: رگرسیون با خطای استاندارد Robust، رگرسیون با داده‌های خوشه‌بندی‌شده، رگرسیون Robust و رگرسیون چندکی.

قبل از اینکه به سراغ این رویکردها برویم، بیایید به یک رگرسیون استاندارد OLS با استفاده از داده‌های شاخص تحصیلی مدارس ابتدایی elemapi2 نگاه کنیم. در این مدل نمرات api00 را با استفاده از متوسط اندازه کلاس acs_k3 و acs_46، درصد معلمان با مدرک کامل full و اندازه مدرسه enroll پیش‌بینی کنیم. ابتدا نگاهی به آمار توصیفی برای این متغیرها می‌کنیم. به مقادیر گمشده acs_k3 و acs_46 توجه داشته باشید.

```
proc means data=elemapi2 mean std max min;
var api00 acs_k3 acs_46 full enroll;
run;
```

The MEANS Procedure

Variable	Label	Mean	Std Dev	Maximum	Minimum
api00	api00	647.6225000	142.2489610	940.0000000	369.0000000
acs_k3	acs_k3	19.1608040	1.3686933	25.0000000	14.0000000
acs_46	acs_46	29.6851385	3.8407840	50.0000000	20.0000000
full	full	84.5500000	14.9497907	100.0000000	37.0000000
enroll	enroll	483.4650000	226.4483847	1570.00	130.0000000

در ادامه مدل رگرسیون برای پیش‌بینی api00 با acs_k3، acs_46، full وenroll را می‌بینیم. تمامی متغیرها به‌جز acs_k3 معنی‌دار هستند.

```
proc reg data=elemapi2;
model api00=acs_k3 acs_46 full enroll;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: api00 api00

Number of Observations Read	400
Number of Observations Used	395
Number of Observations with Missing Values	5

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3071909	767977	61.01	<.0001
Error	390	4909501	12588		
Corrected Total	394	7981410			

Root MSE	112.19832	R-Square	0.3849
Dependent Mean	648.65063	Adj R-Sq	0.3786
Coeff Var	17.29719		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-5.20041	84.95492	-0.06	0.9512
acs_k3	acs_k3	1	6.95438	4.37110	1.59	0.1124
acs_46	acs_46	1	5.96601	1.53105	3.90	0.0001
full	full	1	4.66822	0.41425	11.27	<.0001
enroll	enroll	1	-0.10599	0.02695	-3.93	<.0001

ازآنجایی‌که روش رگرسیون تعاملی است و هنوز دستور quit را صادر نکرده‌ایم می‌توانیم هر دو متغیر اندازه کلاس را آزمون کنیم و می‌بینیم که در آزمون کلی این دو متغیر معنی‌دار هستند.

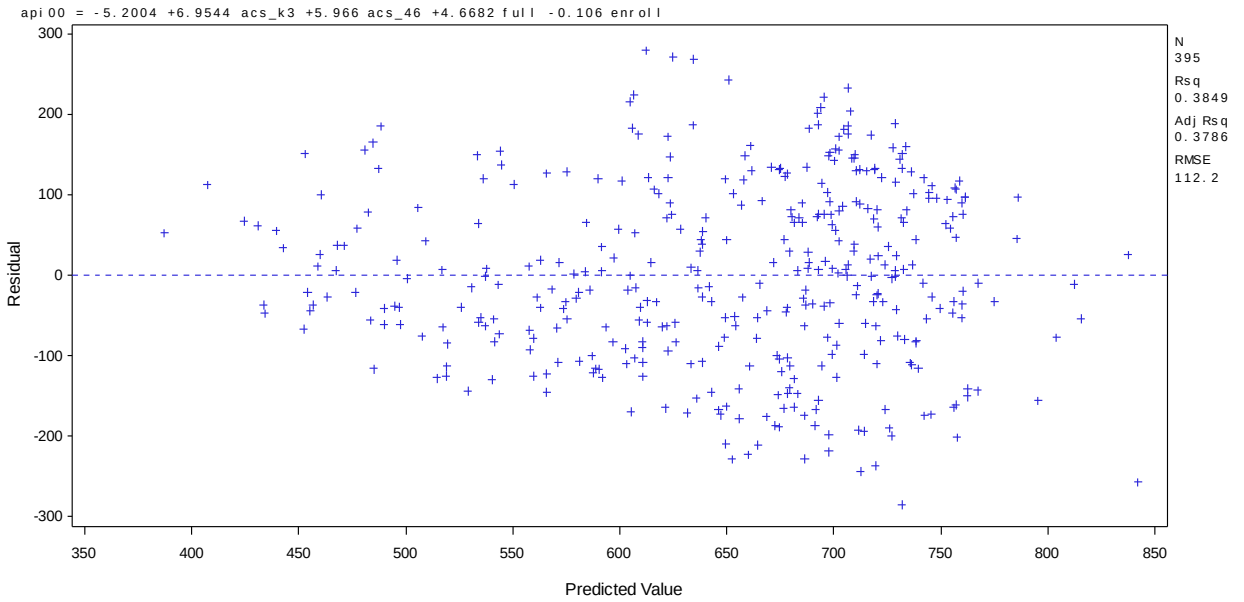
```
test acs_k3=acs_46=0;
run;
```

Test 1 Results for Dependent Variable api00

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	139437	11.08	<.0001
Denominator	390	12588		

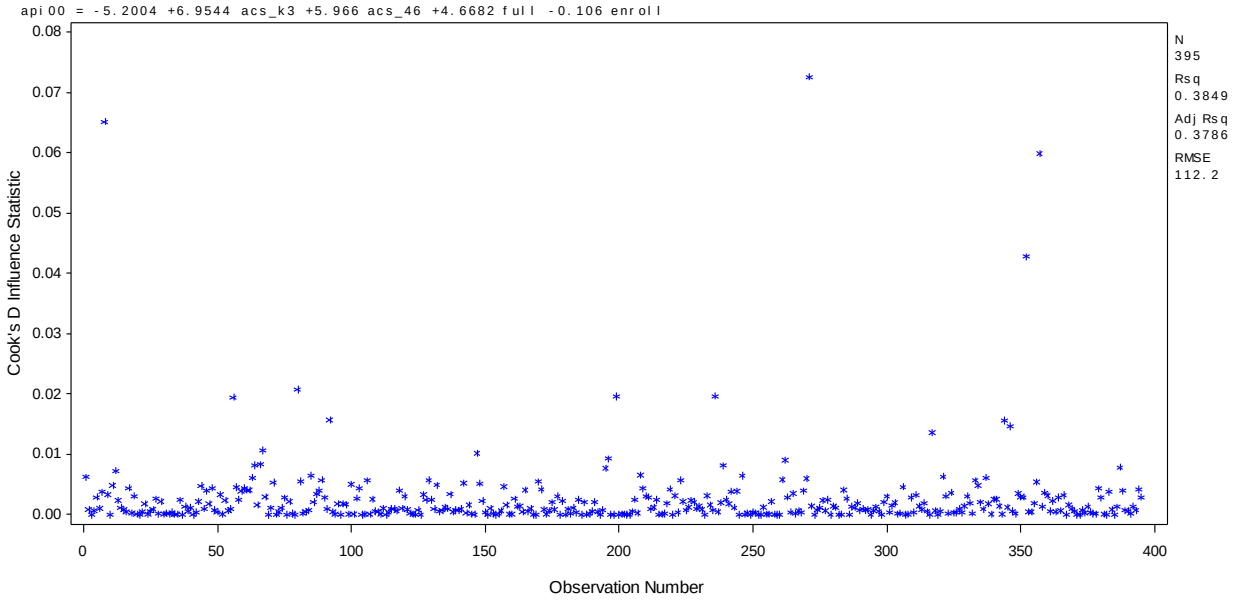
نمودار باقیمانده‌ها در برابر مقادیر پیش‌بینی‌شده از این رگرسیون را رسم می‌کنیم. توجه کنید که الگوی باقیمانده‌ها دقیقاً مطابق انتظار نیست. پراکندگی باقیمانده‌ها در سمت راست نسبت به سمت چپ بیشتر است، جایی که تغییرات باقیمانده‌ها تا حدودی کمتر است و این نشان‌دهنده ناهمگنی است.

```
plot r.*p.;
run;
```



در اینجا نمودار شاخص cook’s D برای این رگرسیون را داریم. می‌بینیم که ۴ نقطه هم در leverage و هم در باقیمانده مقدار بالایی دارند.

```
plot cookd.*obs.;
run;
```



هیچ‌کدام از این نتایج مشکلات اساسی نیست اما نمودار باقیمانده‌ها در برابر مقادیر پیش‌بینی‌شده مقادیر پرت و ناهمگنی در داده‌ها و نمودار شاخص cook's D نقاط مؤثر را در ربع اول را نشان می‌دهد. ممکن است مایل به استفاده از روشی غیر از OLS برای برآورد این مدل باشیم. در چند بخش بعدی به برخی روش‌های رگرسیون Robust را بررسی خواهیم کرد.

رگرسیون با خطای استاندارد Robust

Proc reg در SAS شامل گزینه‌ای به نام acov در ادامه دستور model است که برای برآورد ماتریس کوواریانس نامتقارن برآوردها تحت فرض ناهمگنی است. خطای استاندارد به دست آمده از ماتریس کوواریانس نامتقارن Robust به نظر می‌رسد و می‌تواند با نگرانی‌های جزئی درباره عدم برقراری پیش‌فرض‌ها مثل مسئله نرمالیتی، ناهمگنی یا مشاهداتی که باقیمانده بزرگ دارند، leverage و نفوذپذیری کنار بیاید. برای چنین مشکلات جزئی خطای استاندارد بر اساس acov به‌طور مؤثر می‌تواند با این نگرانی‌ها مقابله کند.

با استفاده از گزینه acov برآورد نقطه ای ضرایب دقیقاً مشابه OLS معمولی هستند اما خطای استاندارد را بر اساس ماتریس کوواریانس نامتقارن محاسبه می‌کنیم. در اینجا همان رگرسیون که در بالا اجرا شد همراه با گزینه acov استفاده می‌کنیم. همچنین از ods در SAS استفاده می‌کنیم تا خروجی برآورد پارامترها تنها با ماتریس کوواریانس نامتقارن باشد. خطای استاندارد Robust را در یک مرحله محاسبه کردیم و با برآورد پارامترها که با استفاده از proc sql ساخته‌ایم ترکیب کردیم و مقادیر t و احتمالات مربوطه را ایجاد کردیم. توجه کنید تغییر در خطای استاندارد و آزمون t (نه در ضرایب) است. در این مثال خاص استفاده از خطای استاندارد Robust هیچ تغییری در نتایج رگرسیون OLS اصلی ایجاد نکرد. همچنین باید ذکر کنیم که خطای استاندارد توانمند برای اصلاح اندازه نمونه تنظیم‌شده است.

```
proc reg data=elemapi2;
model api00=acs_k3 acs_46 full enroll/acov;
ods output Acovest=estcov;
ods output parameterestimates=pest;
run;
quit;
data temp_dm;
set estcov;
drop model dependent;
array a(5) intercept acs_k3 acs_46 full enroll;
array b(5) std1-std5;
b(_n_)=sqrt((395/390)*a(_n_));
std=max(of std1-std5);
keep variable std;
run;
proc sql;
select pest.variable, estimate, stderr, tvalue, probt, std as robust_stderr,
estimate/robust_stderr as tvalue_rb,
(1-probt(abs(estimate/robust_stderr),394))*2 as probt_rb
from pest,temp_dm
where pest.variable=temp_dm. variable;
quit;
```

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|	robust_ stderr	tvalue_rb	probt_rb
Intercept	-5.20041	84.95492	-0.06	0.9512	86.66308	-0.06001	0.95218
acs_k3	6.95438	4.37110	1.59	0.1124	4.620599	1.505082	0.133104
acs_46	5.96601	1.53105	3.90	0.0001	1.573214	3.792246	0.000173
full	4.66822	0.41425	11.27	<.0001	0.414681	11.25737	0
enroll	-0.10599	0.02695	-3.93	<.0001	0.028015	-3.78331	0.000179

استفاده از proc Genmod برای داده‌های خوشه‌بندی‌شده

همان‌طور که در درس ۲ توضیح داده شد رگرسیون OLS فرض می‌کند باقیمانده‌ها مستقل هستند. مجموعه داده elemapi2 حاوی داده‌های مربوط به 400مدرسه از 37 ناحیه مدرسه می‌آیند. بسیار ممکن است که نمرات درون هر منطقه مدرسه مستقل نباشند و این می‌تواند منجر به مستقل نبودن باقیمانده‌ها در منطقه شود. Proc Genmod در SAS برای مدل‌سازی داده‌های همبسته استفاده می‌شود. می‌توانیم از دستور class و repeated استفاده کنیم تا نشان دهیم که مشاهدات در ناحیه‌ها بر اساس dnum خوشه‌بندی‌شده و مشاهدات ممکن است در داخل ناحیه همبسته باشند اما بین ناحیه‌ها مستقل باشند.

```
proc genmod data=elemapi2;
class dnum;
model api00=acs_k3 acs_46 full enroll;
repeated subject=dnum/type=ind;
run;
quit;
```

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-5.2004	119.5172	-239.450	229.0490	-0.04	0.9653
acs_k3	6.9544	6.7726	-6.3196	20.2284	1.03	0.3045
acs_46	5.9660	2.4839	1.0976	10.8344	2.40	0.0163
full	4.6682	0.6904	3.3151	6.0213	6.76	<.0001
enroll	-0.1060	0.0421	-0.1886	-0.0234	-2.51	0.0119

همان‌طور که در رگرسیون با خطای Robust دیدیم برآورد ضرایب مشابه برآورد OLS است، اما خطای استاندارد مشاهدات در ناحیه‌ها را غیرمستقل محاسبه می‌کند. باوجوداینکه خطای استاندارد در این تحلیل بزرگ هستند سه متغیر که در تحلیل OLS معنی‌دار بودند در این تحلیل نیز معنی‌دار هستند.

متوجه می‌شویم که برآورد خطای استاندارد در اینجا متفاوت از نتایج حاصل از استفاده از دستور regress با گزینه خوشه‌بندی در Stata است. این به این دلیل است که Stata بیشتر تنظیمات نمونه را انجام می‌دهد. می‌توانیم برخی از برنامه‌های SAS را برای تعدیل اجرا کنیم. واریانس تعدیل‌شده یک ثابت ضرب‌در واریانس به دست آمده از برآورد خطای استاندارد تجربی است. این ثابت خاص برابر است با (N-1)/(N-k)*M/(M-1)

```
data em;
set elemapi2;
run;
proc genmod data=em;
class dnum;
model api00=acs_k3 acs_46 full enroll;
repeated subject=dnum/type=ind covb;
ods output geercov=gcov;
ods output geeemppest=parms;
run;
quit;
proc sql;
 select count(dnum), count(distinct dnum)into:n, :m
 from em;
 quit;
 proc sql;
 select count(prm1) into:k
 from gcov;
 quit;
 data gcov_ad;
 set gcov;
 array all(*)_numeric_;
 do i=1 to dim(all);
 all(i)=all(i)*((&n-1)/(&n-&k))*(&m/(&m-1));
 if i=_n_ then std_ad=sqrt(all(i));
 end;
 drop i;
 keep std_ad;
 run;
 data all;
 merge parms gcov_ad;
 run;
 proc print data=all noobs;
 run;
```

Parm	Estimate	Stderr	LowerCL	UpperCL	Z	ProbZ	std_ad
Intercept	-5.2004	119.5172	-239.450	229.0490	-0.04	0.9653	121.778
acs_k3	6.9544	6.7726	-6.3196	20.2284	1.03	0.3045	6.901
acs_46	5.9660	2.4839	1.0976	10.8344	2.40	0.0163	2.531
full	4.6682	0.6904	3.3151	6.0213	6.76	<.0001	0.703
enroll	-0.1060	0.0421	-0.1886	-0.0234	-2.51	0.0119	0.043

رگرسیون Robust

در SAS نمی‌توانیم به‌سادگی چند proc را اجرا کنیم تا رگرسیون Robust را با استفاده از تکرار حداقل مربعات باز وزنی اجرا کنیم. به‌منظور انجام رگرسیون Robust باید ماکرو خود را بنویسیم. برای این منظور ATS یک ماکرو به نام sas/webbooks/reg/chapter4/robust_hb.sas نوشته است. این ماکرو ابتدا از وزن Hubert استفاده می‌کند و سپس به وزن دوبل تبدیل می‌شود. به این دلیل است که ماکرو Robust_hb نامیده می‌شود و h و b به ترتیب برای Hubert و biweight هستند. رگرسیون Robust به هر مشاهده با وزن بالاتر وزنی تخصیص می‌دهد که مشاهدات بهتر رفتار کنند. در حقیقت موارد بسیار منحرف آن‌هایی که با cook's D بالاتر از ۱ هستند می‌توانند مجموعه وزن‌های خود را از دست بدهند و بنابراین آن‌ها اصلاً در مدل گنجانده نمی‌شوند.

ماکرو robust_hb.sas از یک ماکرو دیگر به نام sas/webbooks/reg/chapter4/mad.sas برای تولید MAD (میانه انحراف مطلق) در طی روند تکرار استفاده می‌کند. ما هر دو ماکرو را برای انجام تجزیه‌وتحلیل رگرسیون Robust به‌صورت زیر نشان خواهیم داد. توجه کنید که در این تحلیل هر دو ضرایب و خطای استاندارد با رگرسیون OLS اصلی متفاوت است.

```
%include 'S:\software\sas\mad.sas';
%include 'S:\software\sas\robust_hb.sas';
%robust_hb(elemapi2,api00, acs_k3 acs_46 full enroll,
.01,0.00005,10);
```

The REG Procedure
Model: MODEL1
Dependent Variable: api00 api00

Number of Observations Read	400
Number of Observations Used	395
Number of Observations with Missing Values	5

Weight: _w2_

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3053248	763312	73.19	<.0001
Error	390	4067269	10429		
Corrected Total	394	7120516			

Root MSE	102.12196	R-Square	0.4288
Dependent Mean	647.38090	Adj R-Sq	0.4229
Coeff Var	15.77463		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-6.79858	80.45406	-0.08	0.9327
acs_k3	acs_k3	1	6.11031	4.15015	1.47	0.1417
acs_46	acs_46	1	6.25588	1.45280	4.31	<.0001
full	full	1	4.79575	0.39212	12.23	<.0001
enroll	enroll	1	-0.10924	0.02558	-4.27	<.0001

اگر نتایج رگرسیون Robust در بالا با نتایج OLS که قبلاً ارائه‌شده است مقایسه کنید می‌بینید که ضرایب و خطای استاندارد کاملاً مشابه می‌شوند و مقدار t و p-value نیز کاملاً یکسان است. باوجود مشکلات جزئی که در داده‌ها یافتیم وقتی تجزیه‌وتحلیل OLS را انجام دادیم تحلیل رگرسیون Robust نتایج کاملاً مشابهی را نشان می‌دهد که درواقع نشان‌دهنده جزئی بودن این مشکلات است. با توجه به تفاوت قابل‌ملاحظه در نتایج به جستجوی بیشتر برای یافتن علت تفاوت رگرسیون OLS و Robust تمایل داریم و از میان دو نتیجه، نتایج رگرسیون Robust قابل‌اعتمادتر خواهد بود.

بیایید به مقادیر پیش‌بینی‌شده P (fitted)، باقیمانده‌ها r و مقادیر h (hat) leverage نگاهی بیندازیم. ماکرو robust_hb.sas یک مجموعه داده نهایی با مقادیر پیش‌بینی‌شده، باقیمانده‌های خام و مقادیر leverage همراه با مقادیر اصلی به نام _tempout_ تولید می‌کند.

حال اجازه دهید مقادیر پیش‌بینی‌شده متفاوت و وزن‌ها را بررسی کنیم. ابتدا وزن‌های تولیدشده در آخرین تکرار را بر اساس _w2_ مرتب می‌کنیم سپس به ۱۵ مشاهده اول نگاه می‌کنیم. توجه کنید که کوچک‌ترین وزن نزدیک ۱.۵ است اما به‌سرعت در محدوده ۰.۶ قرار می‌گیرد. ۵ مقدار اول به علت مقادیر گمشده پیش‌بینی کننده‌ها گمشده‌اند.

```
proc sort data = _tempout_;
  by descending _w2_;
run;
proc print data = _tempout_ (obs=10);
  var snum api00 p r h _w2_;
run;
```

Obs	snum	api00	p	r	h	_w2_
1	116	513	.	.	.	.
2	3072	763	.	.	.	.
3	3055	590	.	.	.	.
4	4488	521	.	.	.	.
5	4534	445	.	.	.	.
6	637	447	733.426	-286.144	0.003657	0.55684
7	5387	892	611.709	280.464	0.002278	0.57183
8	2267	897	622.062	275.512	0.009887	0.58481
9	65	903	631.930	271.731	0.010198	0.59466
10	3759	585	842.664	-257.473	0.040009	0.63128

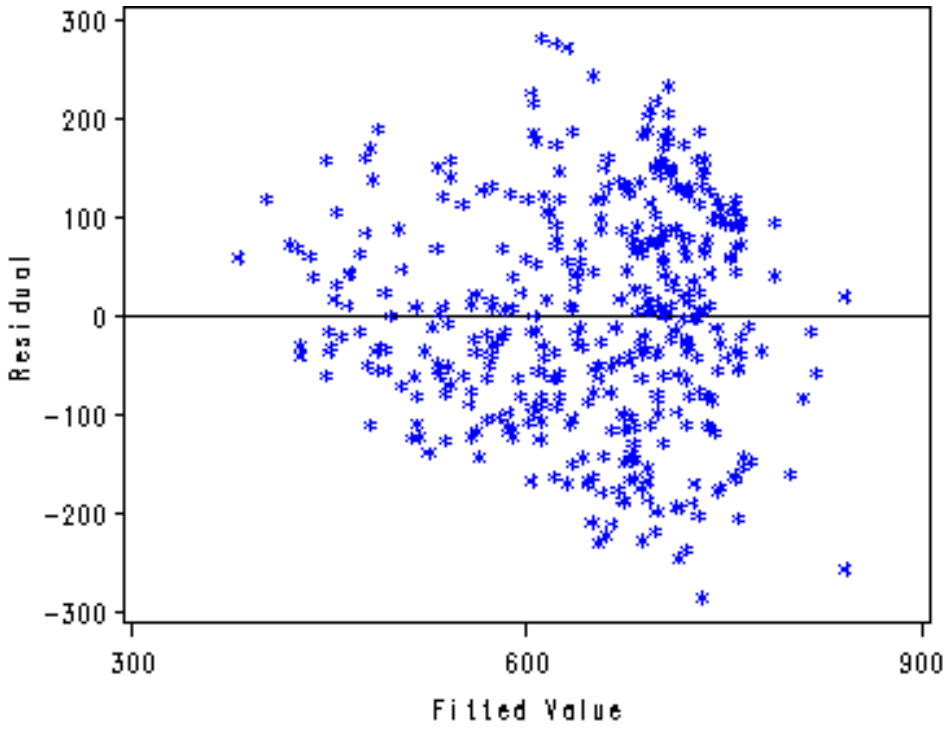
اکنون به ۱۰ مشاهده آخر نگاه می‌کنیم. وزن مشاهدات با snum مقادیر 1678, 4486, 1885 همگی نزدیک به ۱ است وقتی باقیمانده‌ها نسبتاً کوچک هستند. توجه کنید که مشاهدات فوق که کمترین وزن را داشتند بزرگ‌ترین باقیمانده‌ها رادارند (باقیمانده بالاتر از ۲۰۰) و مشاهدات با بالاترین وزن باقیمانده‌های کوچکی داشتند (کمتر از ۳).

```
proc sort data = _tempout_;
  by _w2_;
run;
proc print data = _tempout_ (obs=10);
  var snum api00 p r h _w2_;
run;
```

Obs	snum	api00	p	r	h	_w2_
1	1678	497	497.065	0.18424	0.023137	1.00000
2	4486	706	706.221	0.20151	0.013740	1.00000
3	1885	605	606.066	-0.41421	0.013745	1.00000
4	3535	705	703.929	1.16143	0.004634	0.99999
5	3024	727	729.278	-2.01559	0.010113	0.99997
6	3700	717	719.803	-2.43802	0.007317	0.99996
7	1596	536	533.918	2.77985	0.012143	0.99995
8	181	724	728.068	-3.53209	0.013665	0.99992
9	1949	696	691.898	3.96736	0.020426	0.99990
10	5917	735	731.141	4.03336	0.005831	0.99990

پس از استفاده از ماکرو robust_hb.sas می‌توان از مجموعه داده _tempout_استفاده کنیم تا نمودارهایی برای اهداف تشخیصی رگرسیون ایجاد کنیم. به‌عنوان‌مثال می‌توان نمودار باقیمانده‌ها در مقایسه با مقادیر پیش‌بینی‌شده با خط صفر ایجاد کنیم. این طرح بسیار شبیه به طرح OLS است به‌جز اینکه در طرح OLS همه مشاهدات به‌طور مساوی وزن می‌گیرند اما همان‌طور که مشاهده کردیم مشاهدات با بزرگ‌ترین باقیمانده وزن کمتری می‌گیرند و ازاین‌رو تأثیر کمتری بر نتایج دارند.

```
axis1 order = (-300 to 300 by 100) label=(a=90) minor=none;
axis2 order = (300 to 900 by 300) minor=none;
symbol v=star h=0.8 c=blue;
proc gplot data = _tempout_;
  plot r*p = 1 /haxis=axis2 vaxis=axis1 vref=0;
  label r ="Residual";
  label p ="Fitted Value";
run;
quit;
```



رگرسیون چندکی

به‌طورکلی رگرسیون چندکی و به‌طور خاص رگرسیون میانه می‌تواند به‌عنوان جایگزینی برای رگرسیون robust موردتوجه قرار گیرد. نرم‌افزار SAS رگرسیون چندکی را با استفاده از proc iml اجرا می‌کند. در داخل proc iml روشی به نام LAV نامیده می‌شود و آن یک رگرسیون میانه اجرا می‌کند که ضرایب آن با به حداقل رساندن انحراف مطلق از میانه برآورد می‌شود. البته به‌عنوان یک معیار تمرکز میانه نسبت به میانگین کمتر تحت تأثیر قرار می‌گیرد. البته این به این معنی نیست که رگرسیون میانه یک روش برآورد مقاوم است بلکه درواقع شواهدی وجود دارد که تأثیر مقادیر بالای leverage بر آن را نشان می‌دهد.

در اینجا چگونگی اجرای رگرسیون چندکی با استفاده از proc iml را نشان می‌دهیم. گام اول اطمینان از عدم حضور داده‌های گمشده در مجموعه داده مورداستفاده در proc iml است. در داخل proc iml ابتدا ماتریس‌های لازم برای محاسبات رگرسیون را تولید می‌کنیم و سپس روش LAV را بازخوانی می‌کنیم. پس از فراخوانی LAV می‌توانیم مقادیر پیش‌بینی‌شده و باقیمانده‌ها را محاسبه کنیم. درنهایت مجموعه داده‌ای به نام _temp_ حاوی متغیر وابسته و همه پیشگوها به همراه مقادیر پیش‌بینی‌شده و باقیمانده‌ها ایجاد می‌کنیم.

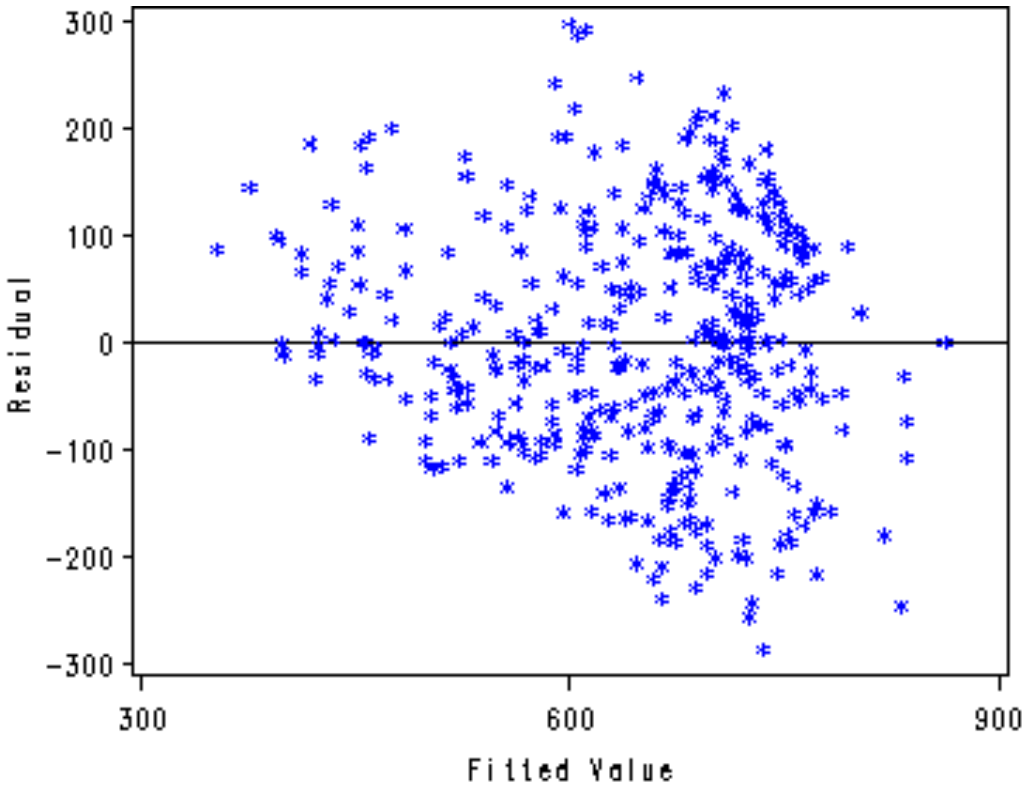
```
data elemapi2;
set elemapi2;
cons=1;
if api00~=. & acs_k3~=. & acs_46~=. & full~=. & enroll~=.;
run;
proc iml ; /*Least absolute values*/
  use elemapi2;
  read all;
  a = cons || acs_k3 || acs_46 || full || enroll;
  b=api00;
```

```
  opt= { . 3  0 1 };
  call lav(rc,xr,a,b,,opt); /*print out the estimates*/

  opt= { . 0  . 1 };
  call lav(rc,xr,a,b,,opt); /*no print out but to create the xr*/

  pred = a*t(xr);
  resid = b - pred;
  create _temp_ var { api00 cons acs_k3 acs_46 full enroll pred resid};
  append;
quit;
```

```
                                LAV (L1) Estimation
                              Start with LS Solution
                        Start Iter: gamma=284.75134813 ActEqn=395
      Iter     N Huber     Act Eqn        Rank       Gamma         L1(x)      F(Gamma)

        31          36           5           5      0.0000     36268.1049     36240.6335

                                  Algorithm converged
                           Objective Function L1(x)= 36268.104941
                             Number Search Directions=       68
                             Number Refactorizations =        2
                             Total Execution Time=       0.0060

                    Necessary and sufficient optimality condition satisfied.


                                  L1 Solution with ASE

Est       17.150502933      1.2690656888      7.2240793844      5.3238408715      -0.124573396
ASE        123.5315456      6.3559394265      2.2262729207       0.602359504      0.0391932684
```

ضریب و خطای استاندارد برای acs_k3 در مقایسه با OLS بسیار متفاوت است. (ضریب ۱.۲ در برابر ۶.۹ و خطای استاندارد ۶.۴ در برابر ۴.۳ است) ضرایب و خطاهای استاندارد برای دیگر متغیرها نیز متفاوت است اما این تفاوت‌ها چشمگیر نیستند. بااین‌وجود نتایج رگرسیون چندکی نشان می‌دهد که، مانند نتایج OLS، تمامی متغیرها به‌جز acs_k3 معنی‌دار هستند. با استفاده از مجموعه داده _temp_ که در بالا ساختیم نمودار باقیمانده‌ها را در برابر مقادیر پیش‌بینی‌شده رسم می‌کنیم.

```
axis1 order=(-300 to 300 by 100) label=(a=90) minor=none;
axis2 order=(300 to 900 by 300) minor=none;
symbol v=star h=0.8 c=blue;
proc gplot data=_tempout_;
plot r*u=1/haxis=axis2 vaxis=axis1 vref=0;
label r="Residual";
label p="Fitted value";
run;
quit;
```



رگرسیون خطی اجباری

بیایید این بخش را با نگاه به مدل رگرسیون با استفاده از مجموعه داده hsb2 آغاز کنیم. فایل hsb2 نمونه‌ای از ۲۰۰ مورد تحصیلات عالی است که متغیرهای id، جنسیت (female)، نژاد، خواندن، نوشتن، علوم اجتماعی و ریاضی را شامل می‌شود. متغیرهای خواندن، نوشتن، علوم اجتماعی و ریاضی به ترتیب نتایج آزمون استانداردشده در خواندن و نوشتن، علوم و مطالعات اجتماعی و ریاضیات است. متغیر جنسیت کد ۱ برای زن و کد ۰ برای مرد تعریف‌شده است. ابتدا با انجام یک رگرسیون OLS نمره socst را از خواندن و نوشتن، ریاضیات و علوم اجتماعی و جنسیت پیش‌بینی می‌کنیم.

```
proc reg data=hsb2;
model socst=read write math science female;
run;
```

```
                The REG Procedure
                  Model: MODEL1
          Dependent Variable: SOCST SOCST

        Number of Observations Read         200
        Number of Observations Used         200
```

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	10948	2189.65226	35.44	<.0001
Error	194	11988	61.79347		
Corrected Total	199	22936			

Root MSE	7.86088	R-Square	0.4773
Dependent Mean	52.40500	Adj R-Sq	0.4639
Coeff Var	15.00025		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	7.32055	3.64859	2.01	0.0462
READ	READ	1	0.37802	0.08066	4.69	<.0001
WRITE	WRITE	1	0.38535	0.08892	4.33	<.0001
MATH	MATH	1	0.12977	0.08932	1.45	0.1479
SCIENCE	SCIENCE	1	-0.03159	0.08150	-0.39	0.6988
FEMALE	FEMALE	1	-0.34790	1.24576	-0.28	0.7803

توجه داشته باشید که ضریب خواندن و نوشتن باهم و همچنین ضریب ریاضیات و علوم باهم شباهت دارند. (درحالی‌که هر دو تفاوت معنی‌داری با صفر ندارند). فرض کنید یک تئوری داریم که خواندن و نوشتن باید ضریب برابر داشته باشند. می‌توانیم با دستور test برابری ضرایب را آزمون کنیم.

```
test read=write;
run;
```

Test 1 Results for Dependent Variable SOCST

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.18350	0.00	0.9566
Denominator	194	61.79347		

نتایج آزمون نشان می‌دهد که تفاوت معنی‌داری بین ضرایب خواندن و نوشتن وجود ندارد. با توجه به برابری ضریب ریاضی و علوم آزمون برابری این دو را نیز انجام می‌دهیم.

```
test math=science;
run;
```

Test 2 Results for Dependent Variable SOCST

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	87.67242	1.42	0.2351
Denominator	194	61.79347		

اکنون این دو آزمون را هم‌زمان باهم انجام می‌دهیم. دستور mtest هم‌زمان آزمون می‌کند که ضریب خواندن برابر نوشتن و ضریب ریاضی برابر با علوم است.

```
mtest math-science,read-write;
run;
```

Multivariate Test 1

Multivariate Statistics and Exact F Statistics

S=1 M=0 N=96

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.99273386	0.71	2	194	0.4929
Pillai's Trace	0.00726614	0.71	2	194	0.4929
Hotelling-Lawley Trace	0.00731933	0.71	2	194	0.4929
Roy's Greatest Root	0.00731933	0.71	2	194	0.4929

توجه داشته باشید که این آزمون دوم دارای ۲ درجه آزادی است زیرا هر دو فرضیه را آزمون می‌کند و این آزمون معنی‌دار نیست نشان می‌دهد که این جفت ضرایب به‌طور معنی‌داری متفاوت از یکدیگر نیستند. درصورتی‌که ضرایب باهم برابر باشند می‌توانیم مدل رگرسیونی را برآورد کنیم. برای مثال با محدودیت برابری ضریب خواندن و نوشتن شروع می‌کنیم. با استفاده از proc reg و دستور restrict این مدل را اجرا می‌کنیم.

```
proc reg data=hsb2;
model socst=read write math science female;
restrict read=write;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: SOCST SOCST

NOTE: Restrictions have been applied to parameter estimates.

Number of Observations Read	200
Number of Observations Used	200

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	10948	2737.01945	44.52	<.0001
Error	195	11988	61.47752		
Corrected Total	199	22936			

Root MSE	7.84076	R-Square	0.4773
Dependent Mean	52.40500	Adj R-Sq	0.4666
Coeff Var	14.96186		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	7.33515	3.62943	2.02	0.0446
READ	READ	1	0.38140	0.05141	7.42	<.0001
WRITE	WRITE	1	0.38140	0.05141	7.42	<.0001
MATH	MATH	1	0.12975	0.08909	1.46	0.1469
SCIENCE	SCIENCE	1	-0.03148	0.08127	-0.39	0.6989
FEMALE	FEMALE	1	-0.32472	1.16790	-0.28	0.7813
RESTRICT		-1	-25.03554	458.25046	-0.05	0.9566*

* Probability computed using beta distribution.

توجه کنید گه ضریب خواندن و نوشتن همراه با خطای استاندارد، آزمون t و غیره یکسان هستند. همچنین دقت کنید که درجه آزادی برای آزمون F مانند مدل OLS برابر ۴ است نه ۵. علت آن است که تنها یک ضریب برای خواندن و نوشتن برآورد می‌شود، مانند یک متغیر تنها که برابر با مجموع مقادیر آن‌ها برآورد می‌شود. همچنین توجه کنید که ریشه MSE برای مدل موردنظر کمی بالاتر است ولی این مقدار جزئی است. علت نیز این است که مدل را مجبور کردیم تا ضریب خواندن و نوشتن در به حداقل رساندن مجموع مربعات خطا یکسان برآورد شود. (ضرایب که موجب کاهش SSE خواهد شد ضرایب مدل بدون مقدار ثابت هستند.)

در ادامه محدودیت دوم را تعریف می‌کنیم و ضریب ریاضی و علوم را باهم برابر قرار می‌دهیم و همراه با محدودیت قبلی مدل را اجرا می‌کنیم.

```
proc reg data=hsb2;
model socst=read write math science female;
restrict read=write , math=science;
run;
```

Dependent Variable: SOCST SOCST

NOTE: Restrictions have been applied to parameter estimates.

Number of Observations Read	200
Number of Observations Used	200

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	10861	3620.17256	58.76	<.0001
Error	196	12076	61.61060		
Corrected Total	199	22936			

Root MSE	7.84924	R-Square	0.4735
Dependent Mean	52.40500	Adj R-Sq	0.4655
Coeff Var	14.97804		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	7.49787	3.63079	2.07	0.0402
READ	READ	1	0.38564	0.05134	7.51	<.0001
WRITE	WRITE	1	0.38564	0.05134	7.51	<.0001
MATH	MATH	1	0.04325	0.05186	0.83	0.4054
SCIENCE	SCIENCE	1	0.04325	0.05186	0.83	0.4054
FEMALE	FEMALE	1	-0.19659	1.16421	-0.17	0.8661
RESTRICT		-1	-15.59490	458.81451	-0.03	0.9730*
RESTRICT		-1	543.08175	455.55392	1.19	0.2342*

* Probability computed using beta distribution.

اکنون ضرایب برای read=write و math=science است و درجه آزادی برای مدل برابر ۳ است. مجدداً ریشه MSE کمی بالاتر از مدل قبلی است اما باید تأکید کنیم که تفاوت جزئی است. اگر درواقع ضریب برای read=write و math=science بود این برآورد ترکیبی ممکن است پایدارتر باشد و بهتر بتوان به دیگر نمونه‌ها تعمیم داد؛ بنابراین اگرچه این برآورد ممکن است سبب افزایش جزئی خطای استاندارد پیش‌بینی‌شده در این نمونه شوند اما ممکن است به جمعیتی که از آن آمده است بهتر باشد.

رگرسیون با داده‌های سانسور و بریده‌شده

تجزیه‌وتحلیل داده‌هایی که حاوی مقادیر سانسور شده و یا بریده‌شده هستند در بسیاری از رشته‌های تحقیقاتی رایج است. با توجه به Hosmer و Lemeshow (1999) مقادیر سانسور شده مقادیری است که به دلیل عوامل تصادفی برای هر موضوع ناقص است. از سوی دیگر مقادیر بریده‌شده به دلیل فرایند انتخاب طرح مطالعه ناقص است.

با بررسی داده‌ها سانسور شده شروع خواهیم کرد.

رگرسیون با مقادیر سانسور شده

در این مثال ما یک متغیر به نام acadindx داریم که ترکیب وزنی از نمرات آزمون استانداردشده و نمرات آکادمیک است. حداکثر نمره ممکن در acadindx برابر ۲۰۰ است اما واضح است که ۱۶ دانش‌آموزی که ۲۰۰ به دست آورده‌اند در توانایی علمی دقیقاً برابر نیستند. به‌عبارت‌دیگر تنوع در توانایی‌های علمی وجود دارد که برای دانش آموزان ۲۰۰ امتیاز در مورد acadindx حساب نمی‌شود. گفته می‌شود متغیر acadindx سانسور شده و به‌خصوص سانسور شده از سمت راست است.

بیایید به مثال نگاه کنیم. با توصیف داده‌ها و آمار توصیفی و همبستگی بین متغیرها شروع می‌کنیم.

```
proc means data=acadindx;
run;
```

```
The MEANS Procedure

Variable      N            Mean         Std Dev         Minimum         Maximum
--------------------------------------------------------------------------------
id          200      100.5000000      57.8791845       1.0000000     200.0000000
female      200        0.5450000       0.4992205               0       1.0000000
reading     200       52.2300000      10.2529368      28.0000000      76.0000000
writing     200       52.7750000       9.4785860      31.0000000      67.0000000
acadindx    200      172.1850000      16.8173987     138.0000000     200.0000000
--------------------------------------------------------------------------------
```

```
proc means data=acadindx n;
where acadindx=200;
run;
```

```
The MEANS Procedure

Variable      N
---------------
id           16
female       16
reading      16
writing      16
acadindx     16
---------------
```

```
proc corr data=acadindx nosimple noprob;
var acadindx female reading writing;
run;
```

```
The CORR Procedure

   4  Variables:    acadindx female   reading  writing

          Pearson Correlation Coefficients, N = 200

              acadindx        female       reading       writing

acadindx       1.00000      -0.08210       0.71309       0.66256

female        -0.08210       1.00000      -0.05308       0.25649

reading        0.71309      -0.05308       1.00000       0.59678

writing        0.66256       0.25649       0.59678       1.00000
```

اکنون بیایید رگرسیون OLS استاندارد را اجرا کنیم و مقادیر پیش‌بینی‌شده را در p1 ذخیره کنیم.

```
proc reg data=acadindx;
model acadindx=female reading writing;
output out=reg1 p=p1;
run;
quit;
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: acadindx

                          Analysis of Variance

                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F

Model                       3         34994          11665     107.40    <.0001
Error                     196         21288      108.61160
Corrected Total           199         56282

Root MSE              10.42169    R-Square     0.6218
Dependent Mean       172.18500    Adj R-Sq     0.6160
Coeff Var              6.05261

                       Parameter Estimates

                     Parameter       Standard
Variable     DF       Estimate          Error    t Value    Pr > |t|

Intercept     1       96.11841        4.48956      21.41      <.0001
female        1       -5.83250        1.58821      -3.67      0.0003
reading       1        0.71842        0.09315       7.71      <.0001
writing       1        0.79057        0.10410       7.59      <.0001
```

Proc lifereg یکی از روش‌های SAS است که می‌تواند برای رگرسیون با داده‌های سانسور شده مورداستفاده قرار گیرد. دستورالعمل شبیه proc reg است. با افزودن متغیری که نشان می‌دهد در مورد مشاهده سانسور شده است؛ بنابراین باید مجموعه داده‌ای با اطلاعات در مورد سانسور کردن ایجاد کنیم.

```
data=tobit_model;
set acadindx;
censor=(acadindx>=200);
run;
proc lifereg data=tobit_model;
model acadindx*censor(1)=female reading writing/d=normal;
output out=reg2 p=p2;
run;
```

```
The LIFEREG Procedure

             Model Information

Data Set                     WORK.TOBIT_MODEL
Dependent Variable                   acadindx
Censoring Variable                     censor
Censoring Value(s)                          1
Number of Observations                    200
Noncensored Values                        184
Right Censored Values                      16
Left Censored Values                        0
Interval Censored Values                    0
Name of Distribution                   Normal
Log Likelihood                   -718.0636168


Algorithm converged.

       Type III Analysis of Effects

                          Wald
Effect       DF    Chi-Square    Pr > ChiSq

female        1       14.0654        0.0002
reading       1       60.8529        <.0001
writing       1       54.1655        <.0001

                  Analysis of Parameter Estimates

                       Standard   95% Confidence     Chi-
Parameter DF Estimate     Error       Limits       Square Pr > ChiSq

Intercept  1  92.7378    4.8034  83.3233 102.1524  372.74     <.0001
female     1  -6.3473    1.6924  -9.6644  -3.0302   14.07     0.0002
reading    1   0.7777    0.0997   0.5823   0.9731   60.85     <.0001
writing    1   0.8111    0.1102   0.5951   1.0271   54.17     <.0001
Scale      1  10.9897    0.5817   9.9067  12.1912
```

بیایید دو مجموعه داده‌ای که ساختیم به‌منظور مقایسه پیش‌بین‌های p1 و p2 باهم ترکیب کنیم.

```
data compare;
merge reg1 reg2;
by id;
run;
proc means data=compare;
var scsdindx p1 p2;
run;
```

```
The MEANS Procedure

Variable      N            Mean         Std Dev         Minimum         Maximum
-------------------------------------------------------------------------------
acadindx    200     172.1850000      16.8173987     138.0000000     200.0000000
p1          200     172.1850000      13.2608696     142.3820725     201.5311070
p2          200     172.7040275      14.0029221     141.2210940     203.8540576
-------------------------------------------------------------------------------
```

این نشان می‌دهد که مقادیر پیش‌بینی‌شده مدل رگرسیون سانسور شده انحراف استاندارد بزرگ‌تر و برد وسیع‌تری دارند. وقتی فهرستی از p1 و p2 را برای همه دانش‌آموزانی که حداکثر ۲۰۰ را در acadindx کسب کرده‌اند نگاه می‌کنیم می‌بینیم که مقادیر پیش‌بینی‌شده در مدل رگرسیون سانسور شده بیش از مقدار پیش‌بینی‌شده OLS است. این پیش‌بینی‌ها برآورد چگونگی تغییرات را نشان می‌دهد اگر مقادیر acadindx می‌توانست از ۲۰۰ بیشتر شود.

```
proc print data=compar;
var acadindx p1=p2;
where acadindx=200;
run;
```

```
Obs    acadindx       p1         p2

 32        200      179.175    179.620
 57        200      192.681    194.329
 68        200      201.531    203.854
 80        200      191.831    193.577
 82        200      188.154    189.563
 88        200      186.573    187.940
 95        200      195.997    198.176
100        200      186.933    188.108
132        200      197.578    199.798
136        200      189.459    191.144
143        200      191.185    192.833
157        200      191.614    193.477
161        200      180.251    181.008
169        200      182.275    183.367
174        200      191.614    193.477
200        200      187.662    189.421
```

Proc lifereg داده‌های راست سانسور و سانسور شده از چپ و سانسور فاصله‌ای را پشتیبانی می‌کند. دو روش دیگر در SAS وجود دارد که رگرسیون سانسور شده را اجرا می‌کند مثل proc qlim.

رگرسیون با داده‌های بریده‌شده

داده‌های بریده‌شده زمانی رخ می‌دهد که برخی از مشاهدات به دلیل مقادیر متغیر در تجزیه‌وتحلیل شامل نمی‌شوند. تجزیه‌وتحلیل داده‌های بریده‌شده را با استفاده از مجموعه داده‌ی acadindx که در بخش قبل استفاده شد نشان خواهیم داد. فرض کنید برای رسیدن به یک برنامه ویژه افتخاری دانش آموزان باید حداقل ۱۶۰ امتیاز از acadindx را کسب کنند؛ بنابراین تمامی مشاهدات را که مقدار acadindx آن‌ها برابر یا کمتر از ۱۶۰ است کنار می‌گذاریم. اکنون همان مدل که در بخش داده‌های سانسور شده اجرا کردیم برآورد می‌کنیم تنها این بار مقدار ۲۰۰ برای acadindx سانسور شده نیست.

```
proc reg data=acadindx;
model acadindx=female reading writing;
where acadindx>160;
run;
quit;
```

```
The REG Procedure
Model: MODEL1
Dependent Variable: acadindx

                          Analysis of Variance

                                   Sum of           Mean
Source                     DF       Squares         Square    F Value    Pr > F

Model                       3     8074.79638     2691.59879      33.01    <.0001
Error                     140         11416       81.54545
Corrected Total           143         19491


Root MSE              9.03025    R-Square     0.4143
Dependent Mean      180.42361    Adj R-Sq     0.4017
Coeff Var             5.00503


                       Parameter Estimates

                    Parameter       Standard
Variable     DF      Estimate          Error    t Value    Pr > |t|

Intercept     1     125.63547        5.89156      21.32      <.0001
female        1      -5.23849        1.61563      -3.24      0.0015
reading       1       0.44111        0.09635       4.58      <.0001
writing       1       0.58733        0.11508       5.10      <.0001
```

واضح است که برآورد ضرایب با توجه به این واقعیت که ۵۳ مشاهده از مجموعه داده کنار گذاشته‌شده‌اند تحریف می‌شود این به محدود کردن برد هر دو متغیر پاسخ و پیشگو کمک می‌کند. برای مثال ضریب write از ۰.۷۹ به ۰.۵۸ کاهش‌یافته است. این به این معنی است که اگر هدف ما یافتن رابطه بین acadindx و متغیرهای پیشگو باشد بریده شدن acadindx در نمونه باعث برآورد اریب می‌شود. روش بهتر برای تحلیل این داده‌ها رگرسیون بریده‌شده است.

با استفاده از proc qlim این کار را انجام می‌دهیم. Proc qlim یک روش آزمایشی است که در نسخه SAS ۸.۱ موجود است. proc qlim(Qualitative and limited dependent variable model) یک مدل متغیر وابسته (و چند متغیره) را که متغیرهای وابسته مقدارهای گسسته را می‌پذیرد یا متغیرهای وابسته تنها در دامنه محدودی از آن‌ها مشاهده می‌شود تحلیل می‌کند.

```
data trunc_model;
set acadindx;
y=.;
if acadindx>160 & acadindx~=. then y=acadindx;
run;
proc qlim data=trunc_model;
model y=female reading writing;
endogenous y~truncated (b=160);
run;
```

```
The QLIM Procedure

                          Summary Statistics of Continuous Responses

                                                                                                N Obs     N
Obs
                          Standard                                Lower       Upper     Lower
Upper
Variable        Mean           Error          Type                Bound       Bound     Bound
Bound

   y        180.4236       11.674837       Truncated             160


            Model Fit Summary

Number of Endogenous Variables            1
Endogenous Variable                       y
Number of Observations                  144
Missing Values                           56
Log Likelihood                 -510.00768
Maximum Absolute Gradient        3.87471E-6
Number of Iterations                     14
AIC                                    1030
Schwarz Criterion                      1045

Algorithm converged.
```

Parameter Estimates

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	110.289206	8.673847	12.72	<.0001
female	-6.099602	1.925245	-3.17	0.0015
reading	0.518179	0.116829	4.44	<.0001
writing	0.766164	0.152620	5.02	<.0001
_Sigma	9.803571	0.721646	13.59	<.0001

ضرایب از proc qlim به نتایج OLSنزدیک‌تر است به‌عنوان‌مثال ضریب نوشتن ۰.۷۷ است مه نزدیک به نتیجه OLS یعنی ۰.۷۹ است. بااین‌حال نتایج هنوز هم تا حدودی در دیگر متغیرها متفاوت است. به‌عنوان‌مثال ضریب خواندن ۰.۵۲ است که نسبت به ۰.۷۲ در OLS با داده‌های نامحدود و ۰.۹۷ OLS با داده‌های محدود بهتراست. درحالی‌که proc qlim ممکن است برآورد یک فایل داده محدودشده را در مقایسه با OLS بهبود بخشد قطعاً هیچ جایگزینی برای تجزیه‌وتحلیل فایل کامل اطلاعات نامحدود ندارد.

رگرسیون با خطای اندازه‌گیری

به‌احتمال‌زیاد به یاد می‌آورید که یکی از پیش‌فرض‌های رگرسیون این است که متغیرهای پیشگو بدون خطای اندازه‌گیری باشند. مشکل این است که خطای اندازه‌گیری در متغیرهای پیشگو منجر به کم برآورد ضرایب رگرسیونی می‌شود.

این بخش درحال‌توسعه است.

مدل‌های رگرسیونی معادله چندگانه

اگر یک مجموعه داده دارای متغیر کافی باشد ممکن است بخواهیم بیش از یک مدل را تخمین بزنیم. برای مثال ممکن است بخواهیم y1 را از x1 و همچنین y2 را از x2 پیش‌بینی کنیم. هرچند که هیچ متغیر مشترکی وجود ندارد این دو مدل مستقل از یکدیگر نیستند زیرا داده‌ها از یک موضوع هستند. این‌یک مثال از رگرسیون معادله چندگانه است که به‌عنوان رگرسیون ظاهراً نامرتبط شناخته می‌شود. می‌توانیم ضرایب را برآورد کنیم و خطای استاندارد را با توجه به خطای مرتبط به مدل به دست آوریم. یک ویژگی مهم از حالت‌های معادله چندگانه این است که می‌توانیم پیشگوها را در معادلات آزمون کنیم. یک مثال دیگر از رگرسیون معادله چندگانه تمایل به پیش‌بینی y1,y2,y3 با x1,x2 است.

این سیستم ۳ معادله به‌عنوان رگرسیون چندمتغیره با متغیرهای پیشگو مشابه برای هر مدل شناخته می‌شود. بازهم توانایی آزمون ضرایب در معادلات مختلف راداریم.

مدل معادلات چندگانه یک افزار قدرتمند برای کیت ابزار تجزیه‌وتحلیل اطلاعات است.

رگرسیون ظاهراً نامرتبط

برای رگرسیون ظاهراً نامرتبط از فایل داده‌های hsb2 استفاده می‌کنیم. این بار به دو مدل رگرسیونی زیر دقت کنید.

Science=math female

Write= read female

این موردی است که خطاهای (باقیمانده‌ها) این دو مدل همبستگی دارند. حتی اگر متغیر پیشگو جنسیت در دو مدل وجود نداشت این موضوع صحیح است. خطاها همبسته خواهند بود؛ زیرا همه مقادیر متغیرها از مجموعه یکسانی از مشاهدات جمع‌آوری‌شده است. این‌یک موقعیت مناسب برای رگرسیون به‌ظاهر نامرتبط با استفاده از proc syslin با گزینه sur است. با استفاده از proc syslin می‌توانیم هر دو مدل را به‌طور هم‌زمان برآورد کنیم درحالی‌که خطای همبسته در همان زمان محاسبه می‌شود منجر به برآورد بسنده ضرایب و خطای استاندارد می‌شود. دستور به‌صورت زیر است:

```
proc syslin data=hsb2 sur;
science: model science=math female;
write: model write=read female;
run;
```

```
                    The SYSLIN Procedure
                Ordinary Least Squares Estimation

                Model                 SCIENCE
                Dependent Variable    SCIENCE
                Label                 SCIENCE


                    Analysis of Variance

                                 Sum of        Mean
Source                 DF       Squares      Square    F Value    Pr > F

Model                   2     8022.016     4011.008      67.96    <.0001
Error                 197     11627.34     59.02203
Corrected Total       199     19649.35


          Root MSE             7.68258    R-Square       0.40826
          Dependent Mean      51.86500    Adj R-Sq       0.40225
          Coeff Var           14.81265
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	18.10509	3.182690	5.69	<.0001	Intercept
MATH	1	0.664059	0.058157	11.42	<.0001	MATH
FEMALE	1	-2.20088	1.091378	-2.02	0.0451	FEMALE

قسمت اول خروجی شامل برآورد OLS برای هر مدل است. اینجا برآورد OLS برای اولین مدل راداریم:

The SYSLIN Procedure
Ordinary Least Squares Estimation

Model	WRITE
Dependent Variable	WRITE
Label	WRITE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7856.321	3928.161	77.21	<.0001
Error	197	10022.55	50.87591		
Corrected Total	199	17878.88			

Root MSE	7.13273	R-Square	0.43942
Dependent Mean	52.77500	Adj R-Sq	0.43373
Coeff Var	13.51537		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	20.22837	2.713756	7.45	<.0001	Intercept
READ	1	0.565887	0.049385	11.46	<.0001	READ
FEMALE	1	5.486894	1.014261	5.41	<.0001	FEMALE

و برآورد OLS برای مدل دوم به‌صورت:

The SYSLIN Procedure
Seemingly Unrelated Regression Estimation

Cross Model Covariance

	SCIENCE	WRITE
SCIENCE	59.0220	7.8780
WRITE	7.8780	50.8759

Cross Model Correlation

	SCIENCE	WRITE
SCIENCE	1.00000	0.14376
WRITE	0.14376	1.00000

Cross Model Inverse Correlation

	SCIENCE	WRITE
SCIENCE	1.02110	-0.14680
WRITE	-0.14680	1.02110

Cross Model Inverse Covariance

	SCIENCE	WRITE
SCIENCE	0.017300	-.002679
WRITE	-.002679	0.020070

System Weighted MSE	0.9981
Degrees of freedom	394
System Weighted R-Square	0.3871

Model	SCIENCE
Dependent Variable	SCIENCE
Label	SCIENCE

Proc syslin با گزینه sur برآورد همبستگی بین خطاهای دو مدل را نیز به ما می‌دهد. خروجی به‌صورت زیر است:

The SYSLIN Procedure
Seemingly Unrelated Regression Estimation

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	20.11558	3.165188	6.36	<.0001	Intercept
MATH	1	0.626086	0.057815	10.83	<.0001	MATH
FEMALE	1	-2.22179	1.091372	-2.04	0.0431	FEMALE

Model	WRITE
Dependent Variable	WRITE
Label	WRITE

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	21.82751	2.699023	8.09	<.0001	Intercept
READ	1	0.535614	0.049095	10.91	<.0001	READ
FEMALE	1	5.453890	1.014245	5.38	<.0001	FEMALE

نهایتاً برآورد رگرسیون ظاهراً نامرتبط برای مدل‌ها راداریم. توجه کنید که هر دو برآورد ضرایب و خطای استاندارد آن‌ها از برآورد مدل OLS نشان داده‌شده در بالا متفاوت است.

The SYSLIN Procedure
Ordinary Least Squares Estimation

Model	SCIENCE
Dependent Variable	SCIENCE
Label	SCIENCE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8022.016	4011.008	67.96	<.0001
Error	197	11627.34	59.02203		
Corrected Total	199	19649.35			

Root MSE	7.68258	R-Square	0.40826
Dependent Mean	51.86500	Adj R-Sq	0.40225
Coeff Var	14.81265		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	18.10509	3.182690	5.69	<.0001	Intercept
MATH	1	0.664059	0.058157	11.42	<.0001	MATH
FEMALE	1	-2.20088	1.091378	-2.02	0.0451	FEMALE

اکنون‌که برآورد مدل‌ها راداریم متغیر پیشگو را آزمون می‌کنیم. آزمون female اطلاعات دو مدل را ترکیب می‌کند. آزمون math و read درواقع برابر با آزمون t بالا است به‌جز اینکه نتایج به‌صورت آزمون F نمایش داده‌شده است.

```
proc syslin data=hsb2 sur;
science: model science=math female;
write: model write=read female;
female: stest science.female=write.female=0;
math: stest science.math=0;
read: stest write.read=0;
run;
```

Test Results

Num DF	Den DF	F Value	Pr > F	Label
2	394	18.52	0.0001	FEMALE

Test Results

Num DF	Den DF	F Value	Pr > F	Label
1	394	117.49	0.0001	MATH

Test Results

Num DF	Den DF	F Value	Pr > F	Label
1	394	119.25	0.0001	READ

اکنون سه مدل را برآورد می‌کنیم که متغیر پیشگو یکسانی در هر مدل که به‌صورت زیر نشان داده استفاده کنیم:

Read= female prog1 prog3

Write= female prog1 prog3

Math= female prog1 prog3

متغیرهای prog1 و prog3 متغیرهای نشانگر برای متغیر prog است. اجازه دهید ابتدا این متغیرها را تولید کنیم؛ قبل از اینکه سه مدل را با استفاده از proc syslin برآورد کنیم.

```
data hsb2;
set hsb2;
prog1=(prog=1);
prog3=(prog=3);
run;
proc syslin data=hsb2 sur;
model1: model read=female prog1 prog3;
model2: model write=female prog1 prog3;
model3: model math=female prog1 prog3;
run;
```

<some output omitted>

The SYSLIN Procedure
Ordinary Least Squares Estimation

Model	MODEL1
Dependent Variable	READ
Label	READ

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	56.82950	1.170562	48.55	<.0001	Intercept
FEMALE	1	-1.20858	1.327672	-0.91	0.3638	FEMALE
prog1	1	-6.42937	1.665893	-3.86	0.0002	
prog3	1	-9.97687	1.606428	-6.21	<.0001	

The SYSLIN Procedure
Ordinary Least Squares Estimation

Model	MODEL2
Dependent Variable	WRITE
Label	WRITE

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	53.62162	1.042019	51.46	<.0001	Intercept
FEMALE	1	4.771211	1.181876	4.04	<.0001	FEMALE
prog1	1	-4.83293	1.482956	-3.26	0.0013	
prog3	1	-9.43807	1.430021	-6.60	<.0001	

The SYSLIN Procedure
Ordinary Least Squares Estimation

Model	MODEL3
Dependent Variable	MATH
Label	MATH

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	57.10551	1.036890	55.07	<.0001	Intercept
FEMALE	1	-0.67377	1.176059	-0.57	0.5674	FEMALE
prog1	1	-6.72394	1.475657	-4.56	<.0001	
prog3	1	-10.3217	1.422983	-7.25	<.0001	

این رگرسیون‌ها برآورد خوب ضرایب و خطاهای استاندارد را ارائه می‌دهند اما این نتایج فرض می‌کند که باقیمانده هر تحلیل کاملاً مستقل از دیگری است. همچنین اگر بخواهیم متغیر جنسیت را آزمون کنیم باید سه با این آزمون را تکرار کنیم و نمی‌توانیم اطلاعات را از هر سه آزمون با یک آزمون کلی ترکیب کنیم.

حالا بیایید خروجی برآورد با استفاده از رگرسیون ظاهراً نامرتبط را ببینیم:

Model	MODEL1
Dependent Variable	READ
Label	READ

The SYSLIN Procedure
Seemingly Unrelated Regression Estimation

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	56.82950	1.170562	48.55	<.0001	Intercept
FEMALE	1	-1.20858	1.327672	-0.91	0.3638	FEMALE
prog1	1	-6.42937	1.665893	-3.86	0.0002	
prog3	1	-9.97687	1.606428	-6.21	<.0001	

Model	MODEL2
Dependent Variable	WRITE
Label	WRITE

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	53.62162	1.042019	51.46	<.0001	Intercept
FEMALE	1	4.771211	1.181876	4.04	<.0001	FEMALE
prog1	1	-4.83293	1.482956	-3.26	0.0013	
prog3	1	-9.43807	1.430021	-6.60	<.0001	

Model	MODEL3
Dependent Variable	MATH
Label	MATH

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variable Label
Intercept	1	57.10551	1.036890	55.07	<.0001	Intercept
FEMALE	1	-0.67377	1.176059	-0.57	0.5674	FEMALE
prog1	1	-6.72394	1.475657	-4.56	<.0001	
prog3	1	-10.3217	1.422983	-7.25	<.0001	

توجه داشته باشید که ضرایب در نتایج OLS و برآورد رگرسیون ظاهراً نامرتبط یکشان هستند بااین‌حال خطای استاندارد درنتیجه همبستگی میان باقیمانده‌ها در معادلات چندگانه فقط کمی متفاوت است.

علاوه برگرفتن خطای استاندارد مناسب‌تر می‌توان اثر متغیرهای پیشگو در معادلات را آزمون کرد. می‌توان فرض کرد که ضریب جنسیت برای هر سه متغیر پاسخ صفر است همان‌طور که در زیر نشان داده‌شده است:

```
proc syslin data=hsb2 sur;
model1: model read=female prog1 prog3;
model2: model write=female prog1 prog3;
model3: model math=female prog1 prog3;
female: stest model1.female=model2.female=model3.female=0;
run;
```

Test Results

Num DF	Den DF	F Value	Pr > F	Label
3	588	11.63	0.0001	FEMALE

همچنین می‌توانیم تنها آزمون برابر صفر بودن ضریب جنسیت در دو معادله read و math را آزمون کنیم.

```
proc syslin data=hsb2 sur;
model1: model read=female prog1 prog3;
model2: model write=female prog1 prog3;
model3: model math=female prog1 prog3;
f1: stest model1.female=model3.female=0;
run;
```

Test Results

Num DF	Den DF	F Value	Pr > F	Label
2	588	0.42	0.6599	F1

و نیز می‌توان آزمون کرد که ضرایب prog1 و prog3 برای هر سه متغیر پاسخ برابر صفر است همان‌طور که در زیر نشان داده‌شده است.

```
proc syslin data=hsb2 sur;
model1: model read=female prog1 prog3;
model2: model write=female prog1 prog3;
model3: model math=female prog1 prog3;
progs: stest model1.prog1=model2.prog1=model3.prog1=0 ,
            model1.prog3=model2.prog3=model3.prog3=0;
run;
```

Test Results

Num DF	Den DF	F Value	Pr > F	Label
6	588	11.83	0.0001	PROGS

رگرسیون چندمتغیره

اکنون با استفاده از رگرسیون چندمتغیره با استفاده از proc reg به دنبال تحلیل مشابهی می‌گردیم که در مثال proc syslin بالا دیدیم، ۳ مدل زیر را برآورد می‌کنیم:

Read=female prog1 prog3

Write= female prog1 prog3

Math= female prog1 prog3

در زیر از proc reg استفاده می‌کنیم تا read,write,math را از متغیرهای پیشگو female, prog1, prog3 پیش‌بینی می‌کنیم. توجه داشته باشید که قسمت بالای خروجی مشابه خروجی sureg است که خلاصه‌ای کلی از هر مدل را برای هر متغیر می‌دهد بااین‌حال نتایج تقریباً متفاوت است و sureg از آزمون chi-square برای آزمون کلی برازش مدل استفاده می‌کند؛ و mureg آزمون F را به کار می‌برد. به نظر می‌رسد بخش پایینی خروجی مشابه خروجی sureg است بااین‌حال هنگامی‌که شما خطای استاندارد را مقایسه می‌کنید می‌بینید که نتایج یکسان نیستند. این خطاهای استاندارد با خطای استاندارد OLS مطابقت دارد بنابراین نتایج زیر همبستگی بین باقیمانده‌ها (به‌عنوان نتایج sureg) را در نظر می‌گیرد.

```
proc reg data=hsb2;
model read write math=female prog1 prog3;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: READ READ

Number of Observations Read	200
Number of Observations Used	200

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3789.28412	1263.09471	14.45	<.0001
Error	196	17130	87.39865		
Corrected Total	199	20919			

Root MSE	9.34872	R-Square	0.1811
Dependent Mean	52.23000	Adj R-Sq	0.1686
Coeff Var	17.89915		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	56.82950	1.17056	48.55	<.0001
FEMALE	FEMALE	1	-1.20858	1.32767	-0.91	0.3638
prog1		1	-6.42937	1.66589	-3.86	0.0002
prog3		1	-9.97687	1.60643	-6.21	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: WRITE WRITE

Number of Observations Read	200
Number of Observations Used	200

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4304.40272	1434.80091	20.72	<.0001
Error	196	13574	69.25751		
Corrected Total	199	17879			

Root MSE	8.32211	R-Square	0.2408
Dependent Mean	52.77500	Adj R-Sq	0.2291
Coeff Var	15.76904		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	53.62162	1.04202	51.46	<.0001
FEMALE	FEMALE	1	4.77121	1.18188	4.04	<.0001
prog1		1	-4.83293	1.48296	-3.26	0.0013
prog3		1	-9.43807	1.43002	-6.60	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: MATH MATH

Number of Observations Read	200
Number of Observations Used	200

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4024.61221	1341.53740	19.56	<.0001
Error	196	13441	68.57746		
Corrected Total	199	17466			

Root MSE	8.28115	R-Square	0.2304
Dependent Mean	52.64500	Adj R-Sq	0.2186
Coeff Var	15.73018		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	57.10551	1.03689	55.07	<.0001
FEMALE	FEMALE	1	-0.67377	1.17606	-0.57	0.5674
prog1		1	-6.72394	1.47566	-4.56	<.0001
prog3		1	-10.32168	1.42298	-7.25	<.0001

اکنون اجازه دهید متغیر female را آزمون کنیم. توجه کنید که این متغیر تنها دریکی از سه معادله معنی‌دار است. با استفاده از دستور mtest بعد از proc reg می‌توانیم به‌طور هم‌زمان female را در سه معادله آزمون کنیم. حدس بزنید چه می‌شود؟ آزمون معنی‌دار است. این نتیجه با آنچه با استفاده از برآورد رگرسیون ظاهراً نامرتبط انجام دادیم سازگار است.

```
female: mtest female=0;
run;
```

The REG Procedure
Model: MODEL1
Multivariate Test: female

Multivariate Statistics and Exact F Statistics

S=1 M=0.5 N=96

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.84892448	11.51	3	194	<.0001
Pillai's Trace	0.15107552	11.51	3	194	<.0001
Hotelling-Lawley Trace	0.17796108	11.51	3	194	<.0001
Roy's Greatest Root	0.17796108	11.51	3	194	<.0001

همچنین می‌توانیم prog1 و prog3 را باهم و به‌صورت مجزا آزمون کنیم. توجه کنید که این آزمون چند متغیره است.

```
prog1: mtest prog1=0;
run;
```

The REG Procedure
Model: MODEL1
Multivariate Test: prog1

Multivariate Statistics and Exact F Statistics

S=1 M=0.5 N=96

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.89429287	7.64	3	194	<.0001
Pillai's Trace	0.10570713	7.64	3	194	<.0001
Hotelling-Lawley Trace	0.11820192	7.64	3	194	<.0001
Roy's Greatest Root	0.11820192	7.64	3	194	<.0001

```
prog3: mtest prog3=0;
run;
```

The REG Procedure
Model: MODEL1
Multivariate Test: prog3

Multivariate Statistics and Exact F Statistics

S=1 M=0.5 N=96

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.75267026	21.25	3	194	<.0001
Pillai's Trace	0.24732974	21.25	3	194	<.0001
Hotelling-Lawley Trace	0.32860304	21.25	3	194	<.0001
Roy's Greatest Root	0.32860304	21.25	3	194	<.0001

```
prog: mtest prog1=prog3=0;
run;
quit;
```

```
                                The REG Procedure
                                  Model: MODEL1
                              Multivariate Test: prog

                    Multivariate Statistics and F Approximations

                                    S=2    M=0    N=96

Statistic                         Value    F Value    Num DF    Den DF    Pr > F

Wilks' Lambda                0.73294667      10.87         6       388    <.0001
Pillai's Trace               0.26859190      10.08         6       390    <.0001
Hotelling-Lawley Trace       0.36225660      11.68         6     256.9    <.0001
Roy's Greatest Root          0.35636617      23.16         3       195    <.0001

          NOTE: F Statistic for Roy's Greatest Root is an upper bound.
                  NOTE: F Statistic for Wilks' Lambda is exact.
```

Proc syslin با گزینه sur و proc reg هر دو شما را قادر به آزمون مدل معادله چندگانه می‌کند درحالی‌که به این واقعیت که معادلات مستقل نیستند توجه می‌کند. Proc syslin با گزینه sur برآورد معادلات با تعدیل معادلات غیرمستقل را ممکن می‌سازد و اجازه می‌دهد معادلاتی را برآورد کنید که الزاماً پیشگو یکسان ندارند. در مقابل proc reg به معادلاتی محدود می‌شود که مجموعه‌ی مشابهی از متغیرهای پیشگو دارند و برآورد آن برای معادلات فردی همانند برآورد OLS است. بااین‌حال proc reg اجازه می‌دهد که شما با آزمون چندمتغیره پیشگوها را به روش سنتی اجرا کنید.

خلاصه

این درس موضوعات متنوعی را پوشش می‌دهد که از رگرسیون حداقل مربعات معمولی فراتر رفته است؛ اما هنوز موضوعات متنوعی که می‌توانستیم پوشش دهیم وجود دارند همچون تجزیه‌وتحلیل بقا، برخورد با داده‌های گمشده، تحلیل داده‌های پانلی و موارد دیگر؛ و برای موضوعاتی که پوشش دادیم آرزو می‌کنیم که می‌توانستیم جزئیات بیشتری ارائه دهیم.. هدف کلی ما در این درس شناخت برخی از تکنیک‌های در دسترس در SAS بود که برای تجزیه‌وتحلیل داده‌هایی که متناسب با پیش‌فرض‌های رگرسیون OLS نیستند استفاده می‌شوند.

Reference https://stats.idre.ucla.edu/?s=regression+with+sas