VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS

Tiesiniai modeliai

Laboratorinis darbas

Atliko: 3 kurso 2 grupės studentai:

Matas Amšiejus

Salvija Račkauskaitė

Sandra Macijauskaitė

Darbo vadovė: doc. dr. Rūta Levulienė

Vilnius, 2021

TURINYS

ĮVADAS

Šiame laboratoriniame darbe analizuosime drabužių siuvyklos duomenis. Taikant tiesinės regresijos modelį bandysime nustatyti kaip priklauso gamyklos produktyvumas nuo įvairių faktorių. Laboratorinio darbo uždavinį įgyvendinti pasitelksime R ir SAS programavimo kalbas.

1. DUOMENYS

1.1.Duomenys

Duomenų rinkinį pasirinkome iš viešai prieinamo duomenų šaltinio „UCI Machine Learning Repository“ (nuoroda šaltiniuose). Duomenyse yra žymimi įvairūs drabužių gamyklos ir jos darbuotojų rodikliai.

1.2.Duomenų aprašymas

Duomenų rinkinį sudarė 1197 stebėjimai su 15 atributų. Modelyje naudosime šiuos:

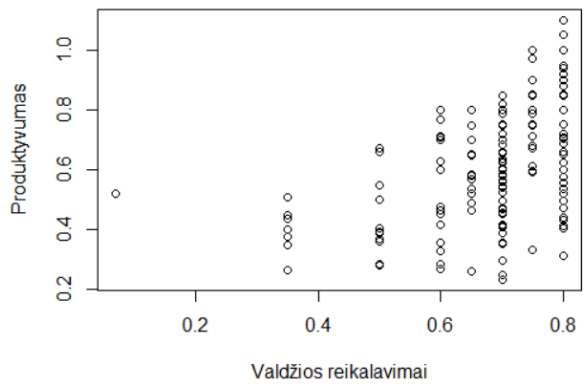
1. Targeted_productivity – gamyklos valdžios nustatyti produktyvumo tikslai ;
2. Smv – užduočiai skiriamas laikas minutėmis;
3. Wip – kiekis trukstamų dalių produktui;
4. Over_time – viršvalandžiai;
5. Incentive – piniginė paskata (BDT (Bangladesh taka) valiuta);
6. No_of_workers – darbuotojų skaičius;
7. Day – savaitės diena;
8. Actual_productivity – tikrasis produktyvumas tą dieną.

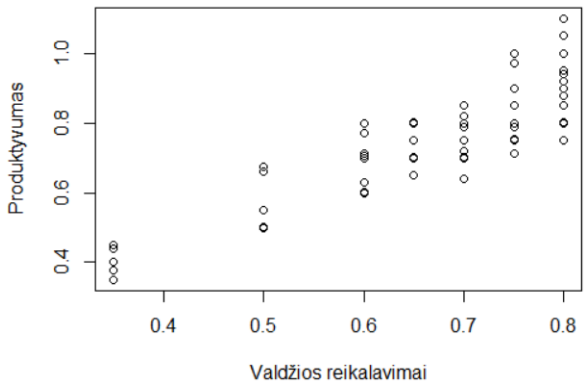
2. ATLIKTAS TYRIMAS

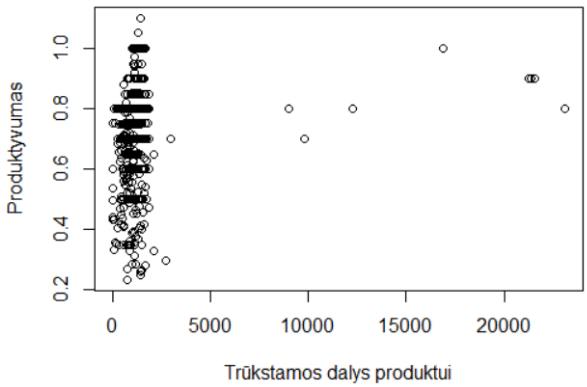
Atlikome daugelio kintamųjų tiesinę regresiją su SAS ir R programavimo kalbomis. Priklausomą kintamąjį pasirenkame Actual_productivity (darbuotojų produktyvumas). Tyrime naudosime reikšmingumo lygmenį alpha = 0,05.

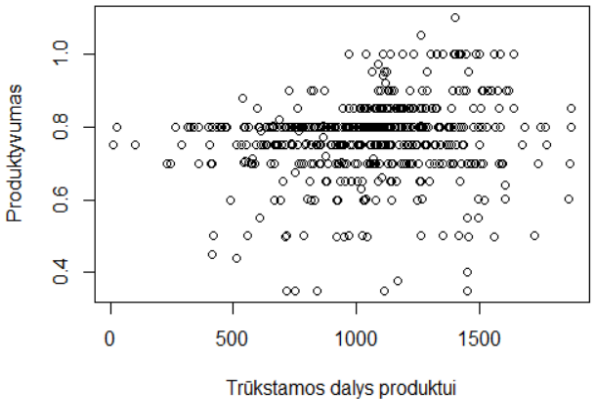
2.1.Bendra tiesinės regresijos eiga

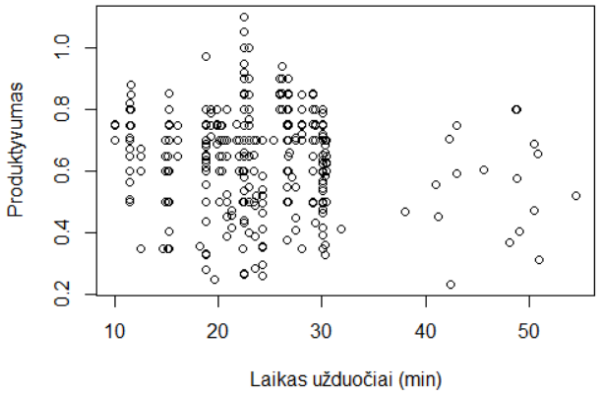
Pirmiausia nuskaitome duomenis ir atrenkame mus dominančius stulpelius. Tada tikriname visų kintamųjų sklaidos diagramas. Pagal jas atsifiltruojame dalį duomenų. Taip pat, kad išvengtume daugiau išskirčių, pasirenkame tik tuos įrašus, kur darbuotojų produktyvumas stipriai nesiskyrė nuo vadovų iškelto tikslo (tenkino bent 90 %).

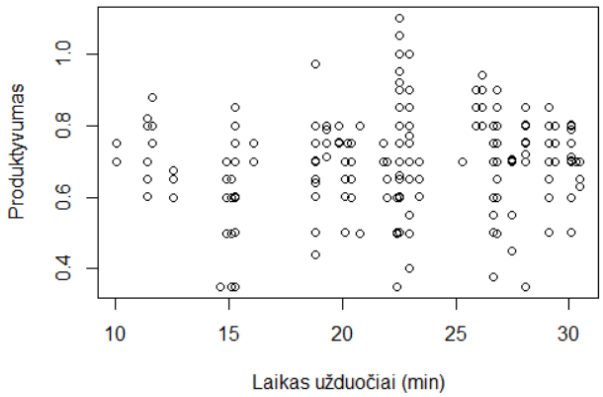

Produktyvumas
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
Valdžios reikalavimai

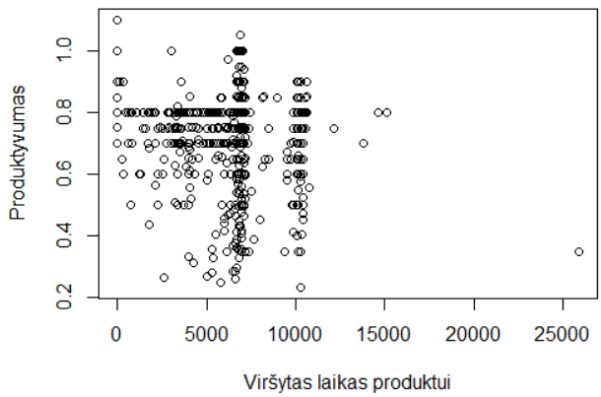

Produktyvumas
0.4
0.6
0.8
1.0
0.4
0.5
0.6
0.7
0.8
Valdžios reikalavimai

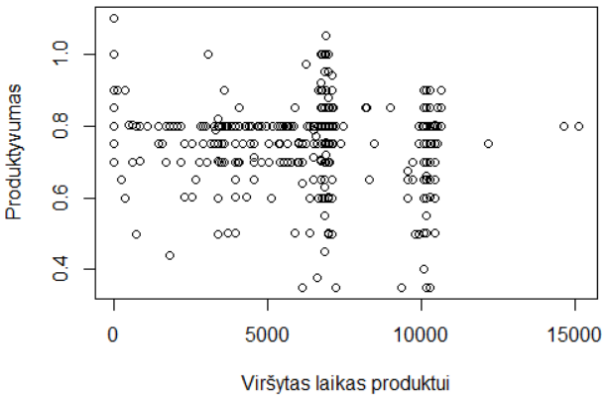

Produktyvumas
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000
20000
Trūkstamos dalys produktui

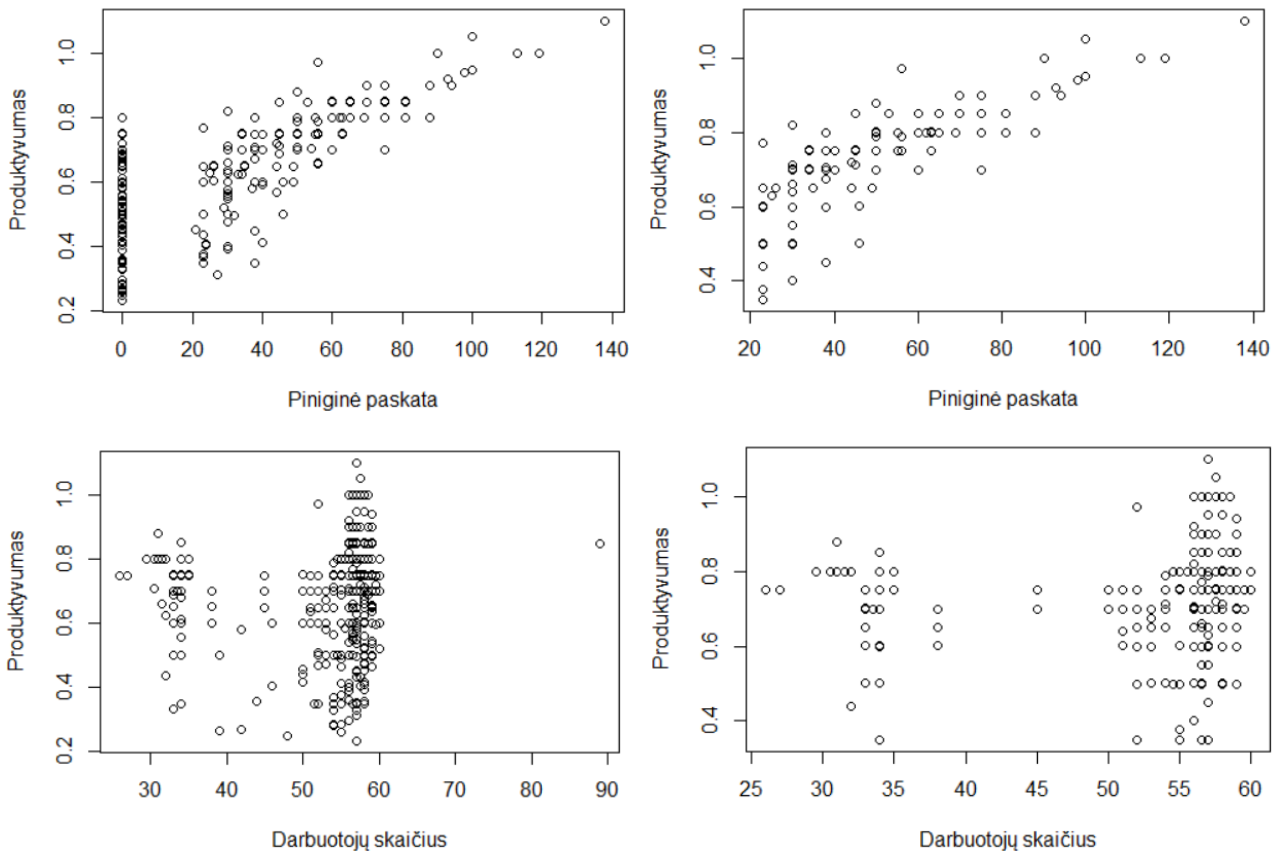

Produktyvumas
0.4
0.6
0.8
1.0
0
500
1000
1500
Trūkstamos dalys produktui


Produktyvumas
0.2
0.4
0.6
0.8
1.0
10
20
30
40
50
Laikas užduočiai (min)


Produktyvumas
0.4
0.6
0.8
1.0
10
15
20
25
30
Laikas užduočiai (min)


Produktyvumas
0.2
0.4
0.6
0.8
1.0
0
5000
10000
15000
20000
25000
Viršytas laikas produktui


Produktyvumas
0.4
0.6
0.8
1.0
0
5000
10000
15000
Viršytas laikas produktui

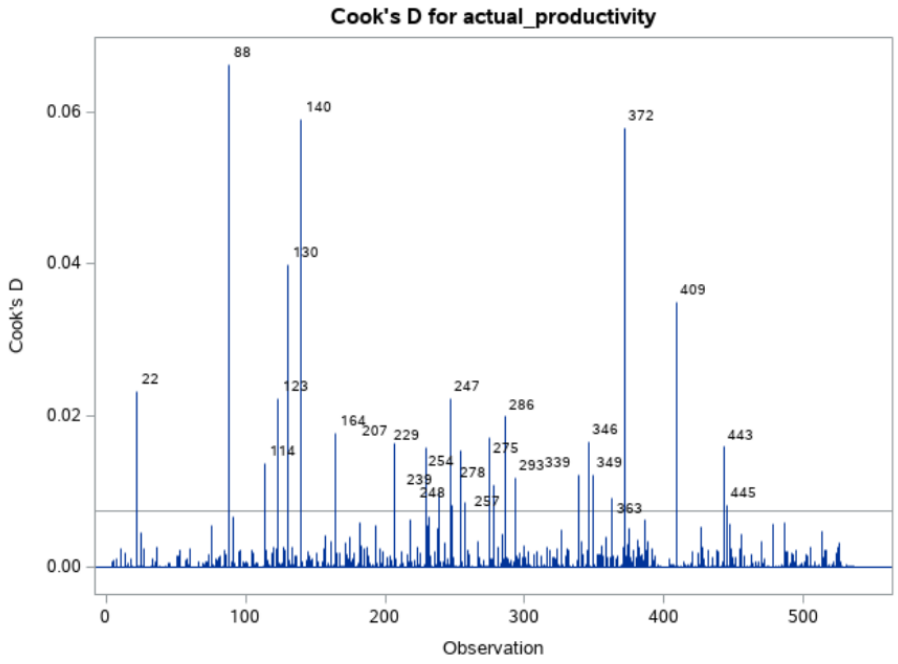


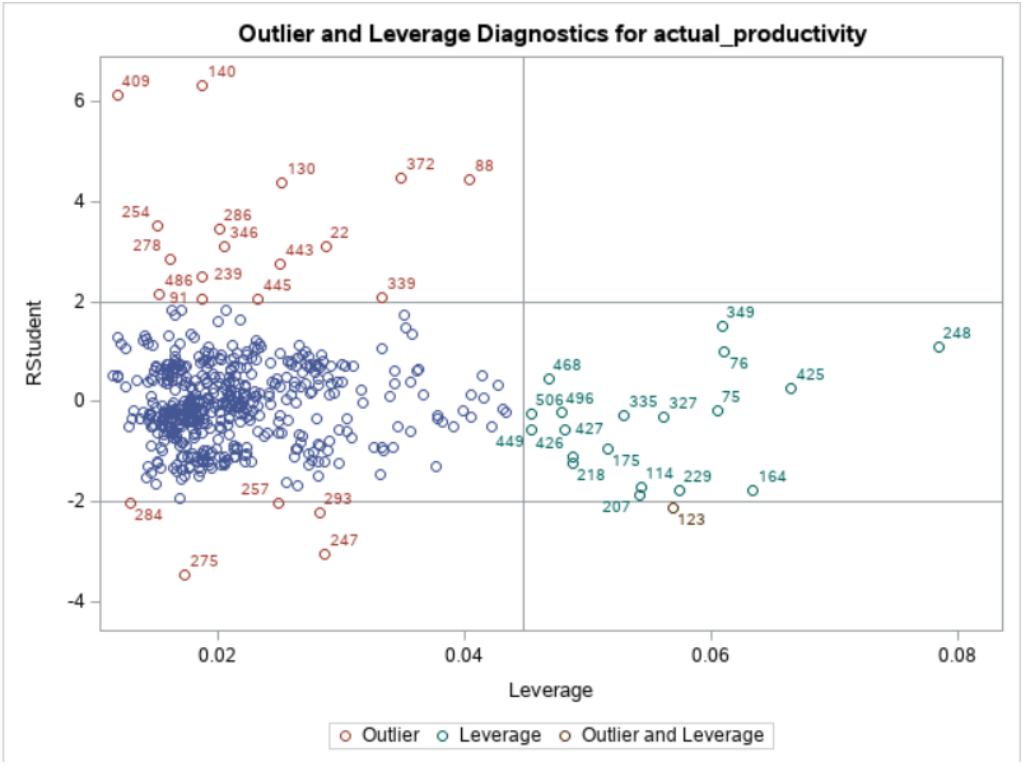
Tikriname, kaip priklausomas kintamasis (darbuotojų produktyvumas) koreliuoja su kovariantėmis.

Pearson Correlation Coefficients, N = 536 Prob > |r| under H0: Rho=0							
	actual_productivity	targeted_productivity	smv	incentive	wip	over_time	no_of_workers
actual_productivity	1.00000	0.85139 <.0001	0.02225 0.6073	0.82689 <.0001	0.17035 <.0001	-0.04349 0.3149	0.07628 0.0777
targeted_productivity	0.85139 <.0001	1.00000	-0.05425 0.2098	0.52616 <.0001	-0.01047 0.8090	-0.10338 0.0167	-0.08960 0.0381
smv	0.02225 0.6073	-0.05425 0.2098	1.00000	0.06576 0.1283	0.02102 0.6273	0.29594 <.0001	0.69916 <.0001
incentive	0.82689 <.0001	0.52616 <.0001	0.06576 0.1283	1.00000	0.28742 <.0001	0.08898 0.0395	0.19152 <.0001
wip	0.17035 <.0001	-0.01047 0.8090	0.02102 0.6273	0.28742 <.0001	1.00000	0.17337 <.0001	0.07974 0.0651
over_time	-0.04349 0.3149	-0.10338 0.0167	0.29594 <.0001	0.08898 0.0395	0.17337 <.0001	1.00000	0.37285 <.0001
no_of_workers	0.07628 0.0777	-0.08960 0.0381	0.69916 <.0001	0.19152 <.0001	0.07974 0.0651	0.37285 <.0001	1.00000

Matome, kad koreliacija pakankamai stipri su parinktomis kovariantėmis, todėl galime bandyti nustatyti tiesinės regresijos modelį.

Sukuriame tiesinės regresijos modelį. Tikriname išskirtis pagal Kuko (Cook‘s D) ir stjudentizuotų paklaidų (R student) kriterijus.





Matome, kad pagal Kuko kriterijų išskirčių nėra. Tačiau pagal Rstudent jų turime nemažai. Šaliname išskirtis. Pastaba: šalinsime ne visas iš karto, o atsižvelgiant į tai, kaip modelis kis po individualios išskirties pašalinimo.

Pašalinus visas išskirtis, gauname labai gerą R square reikšmę.

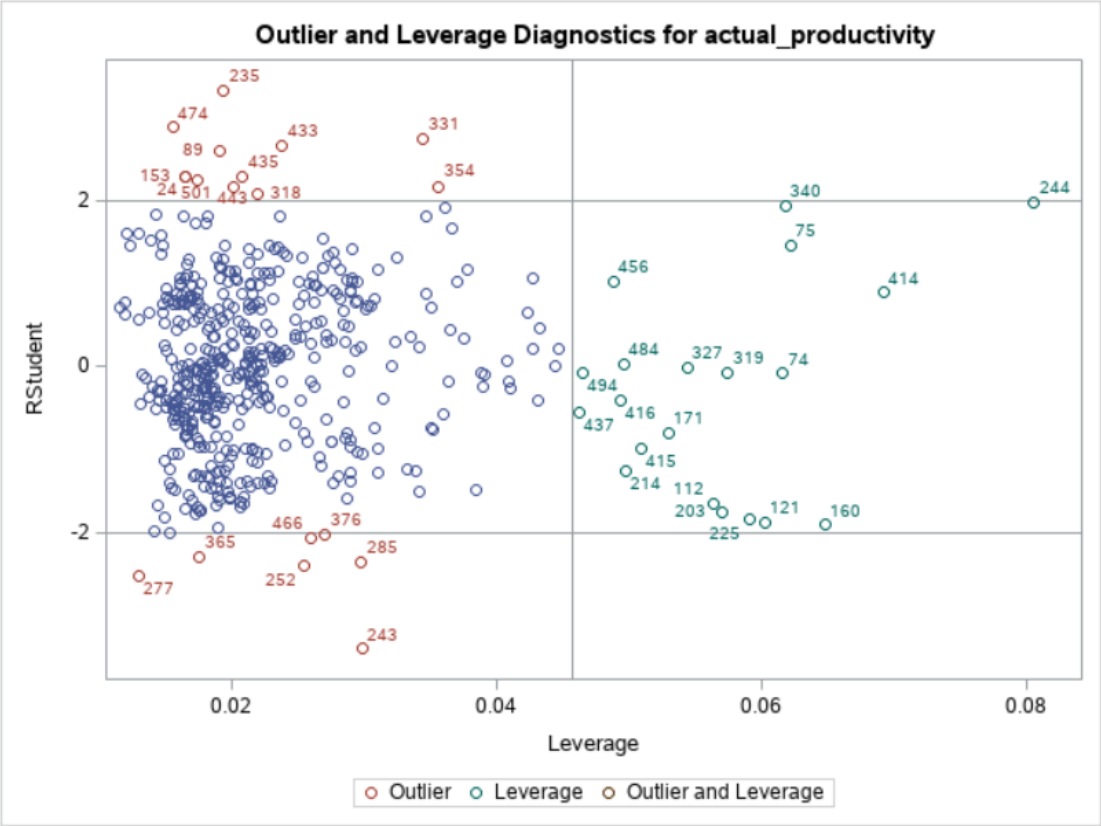
The REG Procedure
Model: MODEL1
Dependent Variable: actual_productivity

Number of Observations Read	524
Number of Observations Used	524

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	6.18686	0.56244	972.00	<.0001
Error	512	0.29627	0.00057865		
Corrected Total	523	6.48312			

Root MSE	0.02406	R-Square	0.9543
Dependent Mean	0.76858	Adj R-Sq	0.9533
Coeff Var	3.12981		



Tikriname, kad paklaidos pasiskirsčiusios pagal normalųjį skirstinį. Tam naudosime Shapiro – Wilk normalumo testą.

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.995692	Pr < W	0.1597
Kolmogorov-Smirnov	D	0.038551	Pr > D	0.0576
Cramer-von Mises	W-Sq	0.109628	Pr > W-Sq	0.0870
Anderson-Darling	A-Sq	0.761857	Pr > A-Sq	0.0477

Gauname, kad p reikšmė daugiau už reikšmingumo lygmenį alpha = 0,05, todėl nulinės hipotezės atmesti negalime. Paklaidos tenkina normalumo prielaidą.

Dabar tikrinsime homoskedastiškumo prielaidą (paklaidų dispersijos lygios). Tam naudosime Breusch – Pagan homoskedastiškumo testą.

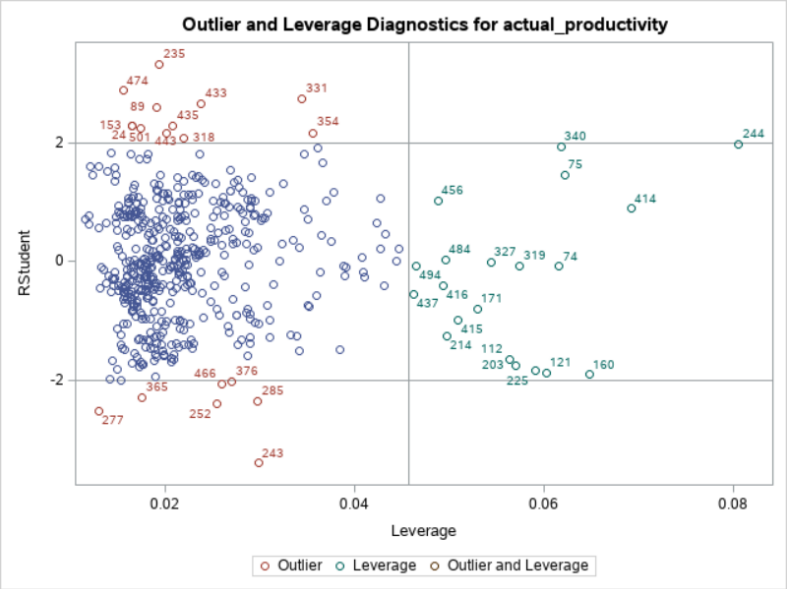
Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
actual_productivity	White's Test	131.1	62	<.0001	Cross of all vars
	Breusch-Pagan	45.24	11	<.0001	1, targeted_productivity, smv, wip, over_time, incentive, no_of_workers, day1, day2, day3, day4, day5

Gauname, kad p reikšmė yra mažiau už reikšmingumo lygmenį, todėl nulinę hipotezę atmetame. Paklaidų dispersijos yra nevienodos (heteroskedastiškos), todėl naudosime HC_0 korekciją. Gauname naujas pataisytas standartines paklaidas bei stulpelių reikšmingumo p reikšmes.

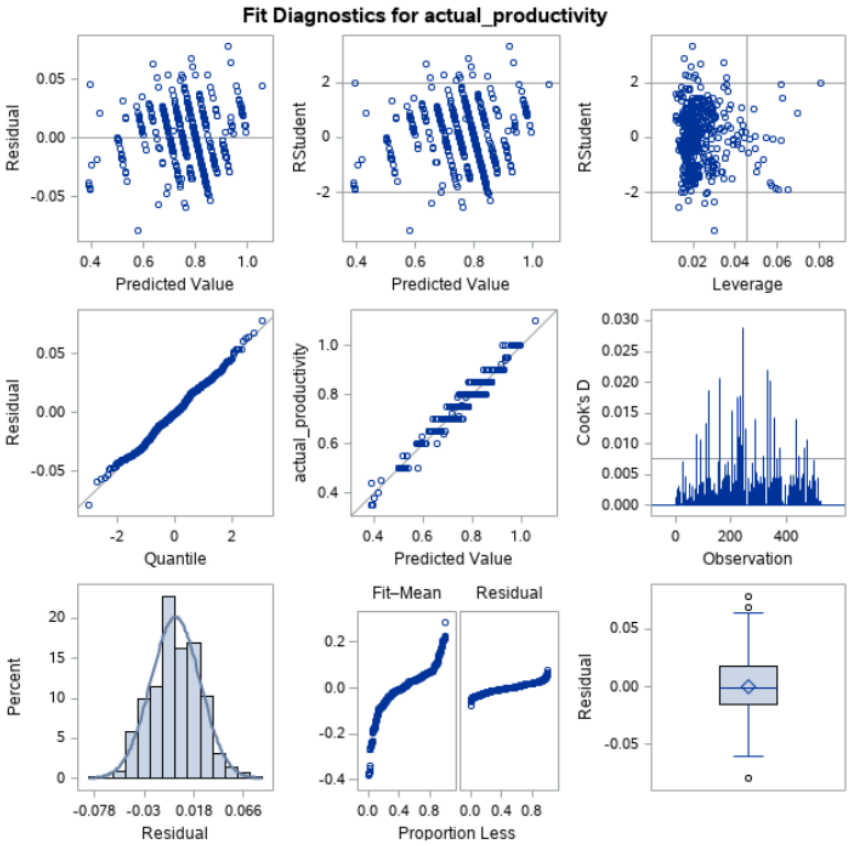
Root MSE	0.02406	R-Square	0.9543
Dependent Mean	0.76858	Adj R-Sq	0.9533
Coeff Var	3.12981		

Parameter Estimates														
						Heteroscedasticity Consistent							Heteroscedasticity Consistent 95% Confidence Limits	
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Partial Corr Type II	Variance Inflation	95% Confidence Limits		
Intercept	1	0.04672	0.01297	3.60	0.0003	0.01542	3.03	0.0026	0	.	0	0.02123	0.07220	0.01643	0.07701
targeted_productivity	1	0.74896	0.01461	51.28	<.0001	0.01860	40.27	<.0001	0.60053	0.83703	1.53651	0.72027	0.77765	0.71242	0.78550
smv	1	0.00038473	0.00025694	1.50	0.1349	0.00023090	1.67	0.0963	0.02013	0.00436	2.02426	-0.00012006	0.00088951	-0.00006891	0.00083836
wip	1	0.00001644	0.00000356	4.62	<.0001	0.00000349	4.71	<.0001	0.04722	0.04006	1.16948	0.00000945	0.00002343	0.00000958	0.00002330
over_time	1	-0.00000204	4.068033E-7	-5.01	<.0001	4.172976E-7	-4.88	<.0001	-0.05238	0.04669	1.22569	-0.00000284	-0.00000124	-0.00000286	-0.00000122
incentive	1	0.00261	0.00006442	40.57	<.0001	0.00007109	36.76	<.0001	0.50654	0.76272	1.74671	0.00249	0.00274	0.00247	0.00275
no_of_workers	1	0.00034689	0.00016225	2.14	0.0330	0.00014176	2.45	0.0147	0.03050	0.00885	2.27995	0.00002813	0.00066566	0.00006840	0.00062539
day1	1	0.00191	0.00377	0.51	0.6134	0.00379	0.50	0.6153	0.00629	0.00049880	1.73600	-0.00551	0.00932	-0.00555	0.00936
day2	1	0.00011521	0.00364	0.03	0.9748	0.00350	0.03	0.9737	0.00040387	0.00000195	1.82744	-0.00704	0.00727	-0.00676	0.00699
day3	1	0.00079117	0.00369	0.21	0.8305	0.00388	0.20	0.8384	0.00271	0.00008954	1.78959	-0.00647	0.00805	-0.00683	0.00841
day4	1	-0.00207	0.00375	-0.55	0.5810	0.00363	-0.57	0.5682	-0.00692	0.00059523	1.76183	-0.00943	0.00529	-0.00919	0.00505
day5	1	-0.00671	0.00378	-1.77	0.0766	0.00361	-1.86	0.0637	-0.02202	0.00611	1.72542	-0.01413	0.00071981	-0.01380	0.00038252

Iš lentelės matome, kad ne visos kovariantės yra reikšmingos. Svarbiausios yra targeted_productivity, wip, over_time ir incentive. Modelio R square reikšmė gaunasi labai gera (apie 95 %).



Pagal Rstudent kriterijų išskirčių nebeliko.



Tobuliname modelį. Atrenkame tik reikšmingas kovariantes. Tam naudosime pažingsninę regresiją (reikšmingumo lygmuo 0,05).

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	targeted_productivity		1	0.7486	0.7486	2297.00	1554.13	<.0001
2	incentive		2	0.2004	0.9490	53.5660	2046.78	<.0001
3	wip		3	0.0014	0.9504	39.8520	14.70	0.0001
4	over_time		4	0.0013	0.9517	27.0843	14.17	0.0002
5	no_of_workers		5	0.0018	0.9535	9.1568	19.81	<.0001
6	day5		6	0.0005	0.9540	5.5382	5.63	0.0180

Nustatėme, kad smv yra nereikšminga. Išmetus smv matome, kad savaitės dienos irgi tampa nereikšmingos, todėl pašaliname jas iš modelio.

Parameter Estimates														
						Heteroscedasticity Consistent								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Partial Corr Type II	Variance Inflation	95% Confidence Limits		Heteroscedasticity Consistent 95% Confidence Limits	
Intercept	1	0.04609	0.01298	3.55	0.0004	0.01567	2.94	0.0034	0	.	0	0.02058	0.07159	0.01529	0.07688
targeted_productivity	1	0.75049	0.01459	51.45	<.0001	0.01891	39.68	<.0001	0.60176	0.83766	1.52899	0.72183	0.77915	0.71334	0.78765
wip	1	0.00001629	0.00000356	4.58	<.0001	0.00000354	4.60	<.0001	0.04679	0.03923	1.16856	0.00000930	0.00002329	0.00000933	0.00002325
over_time	1	-0.00000199	4.062458E-7	-4.91	<.0001	4.141723E-7	-4.81	<.0001	-0.05125	0.04483	1.21938	-0.00000279	-0.00000120	-0.00000281	-0.00000118
incentive	1	0.00260	0.00006411	40.60	<.0001	0.00007066	36.84	<.0001	0.50450	0.76265	1.72579	0.00248	0.00273	0.00246	0.00274
no_of_workers	1	0.00050892	0.00012104	4.20	<.0001	0.00011185	4.55	<.0001	0.04474	0.03331	1.26587	0.00027112	0.00074673	0.00028918	0.00072867
day1	1	0.00201	0.00378	0.53	0.5943	0.00381	0.53	0.5976	0.00664	0.00055329	1.73540	-0.00541	0.00943	-0.00547	0.00950
day2	1	0.00022324	0.00365	0.06	0.9512	0.00349	0.06	0.9490	0.00078258	0.00000730	1.82672	-0.00694	0.00739	-0.00663	0.00708
day3	1	0.00093346	0.00370	0.25	0.8008	0.00387	0.24	0.8095	0.00319	0.00012417	1.78841	-0.00633	0.00820	-0.00667	0.00854
day4	1	-0.00205	0.00375	-0.55	0.5851	0.00361	-0.57	0.5706	-0.00686	0.00058155	1.76181	-0.00942	0.00532	-0.00915	0.00505
day5	1	-0.00668	0.00379	-1.76	0.0782	0.00361	-1.85	0.0652	-0.02192	0.00603	1.72537	-0.01412	0.00075721	-0.01378	0.00042191

Pašalinus dienų faktorių (su visais pseudokintamaisiais) turime galutinį tiesinės regresijos modelį. VIF niekur nesiekia 4, todėl daugiau nieko nekeičiame. Galutinė Adjusted R Square reikšmė yra 95,3 % (tokia dalis duomenų paaiškinama tiesinės regresijos modeliu).

The REG Procedure
Model: MODEL1
Dependent Variable: actual_productivity

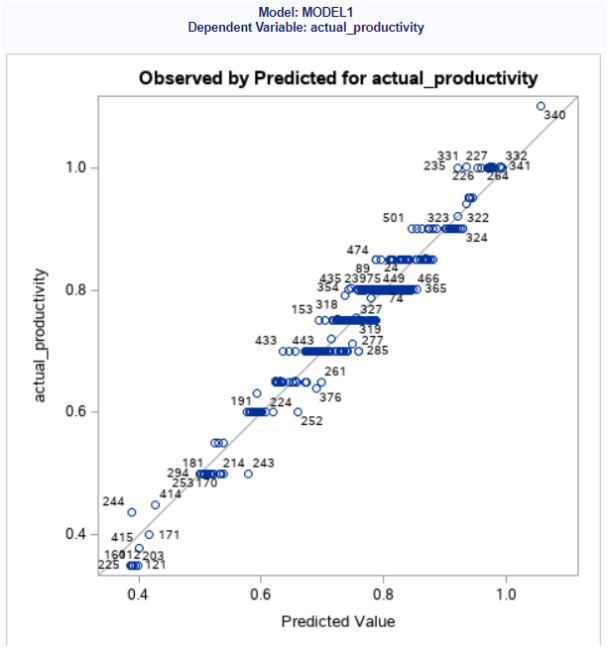
Number of Observations Read	524
Number of Observations Used	524

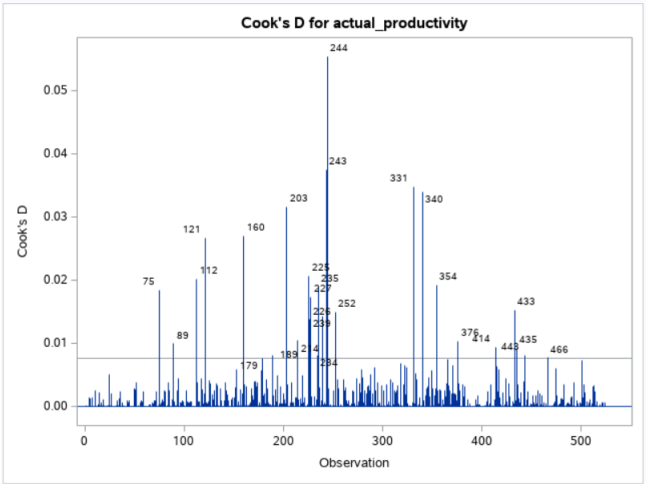
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6.18156	1.23631	2123.62	<.0001
Error	518	0.30157	0.00058217		
Corrected Total	523	6.48312			

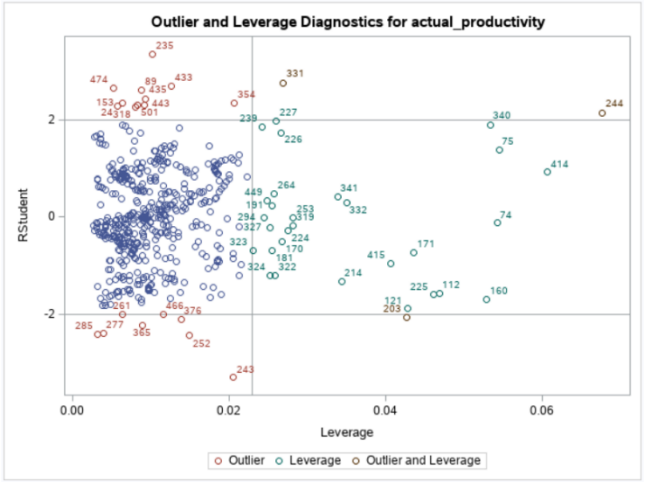
Root MSE	0.02413	R-Square	0.9535
Dependent Mean	0.76858	Adj R-Sq	0.9530
Coeff Var	3.13933		

Parameter Estimates															
						Heteroscedasticity Consistent									
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Partial Corr Type II	Variance Inflation	95% Confidence Limits		Heteroscedasticity Consistent 95% Confidence Limits	
Intercept	1	0.04320	0.01263	3.42	0.0007	0.01539	2.81	0.0052	0	.	0	0.01840	0.06801	0.01296	0.07344
targeted_productivity	1	0.75301	0.01454	51.77	<.0001	0.01867	40.33	<.0001	0.60378	0.83804	1.51457	0.72444	0.78158	0.71633	0.78969
wip	1	0.00001675	0.00000355	4.72	<.0001	0.00000354	4.73	<.0001	0.04810	0.04119	1.15801	0.00000977	0.00002372	0.00000978	0.00002371
over_time	1	-0.00000208	4.042118E-7	-5.15	<.0001	4.16823E-7	-4.99	<.0001	-0.05349	0.04865	1.20279	-0.00000287	-0.00000129	-0.00000290	-0.00000126
incentive	1	0.00258	0.00006342	40.71	<.0001	0.00006915	37.33	<.0001	0.50039	0.76188	1.68242	0.00246	0.00271	0.00245	0.00272
no_of_workers	1	0.00053556	0.00012034	4.45	<.0001	0.00011088	4.83	<.0001	0.04709	0.03683	1.24656	0.00029915	0.00077197	0.00031774	0.00075338

Gauti galutiniai parametrų įvertiniai: $\beta_0 \approx 0{,}0432$, $\beta_1 \approx 0{,}753$, $\beta_2 \approx 0{,}000017$, $\beta_3 \approx -0{,}000002$, $\beta_4 \approx 0{,}00258$, $\beta_5 \approx 0{,}00054$. Didžiausią įtaką gamyklos produktyvumui daro vadovų tikslai ir piniginė paskata.







IŠVADOS

Atlikus pilną regresinę analizę sukūrėme gana tikslų modelį, pritaikytą nustatyti gamyklos produktyvumą pagal svarbiausias kovariantes. Didžiausią įtaką produktyvumui turėjo valdžios iškelti reikalavimai bei piniginė paskata. Įtakos nedarė savaitės diena ir užduočiai skiriamas laikas minutėmis.

ŠALTINIAI

[1] „UCI Machine Learning Repository“ tinklapis. Tema: Productivity Prediction of Garment Employees. Prieiga per internetą: https://archive.ics.uci.edu/ml/datasets/Productivity+Prediction+of+Garment+Employees

[2] Heteroskedastiškumo pavyzdžiai su R. Prieiga per internetą: Dealing with heteroskedasticity; regression with robust standard errors using R (brodrigues.co)

[3] „SAS Help Center“ tinklapis. Prieiga per internetą: SAS Help Center: SAS Help Center: Welcome