

MAAŞ TAHMIN MODELİ

Çoklu Doğrusal Regresyon

İrfan Yakut





Veri Seti Hikayesi

Bu veri seti orijinal olarak Carnegie Mellon Üniversitesi'nde bulunan StatLib kütüphanesinden alınmıştır.

Veri seti 1988 ASA Grafik Bölümü Poster Oturumu'nda kullanılan verilerin bir parçasıdır.

Maaş verileri orijinal olarak Sports Illustrated, 20 Nisan 1987'den alınmıştır.

1986 ve kariyer istatistikleri, Collier Books, Macmillan Publishing Company, New York tarafından yayınlanan 1987 Beyzbol Ansiklopedisi Güncellemesinden elde edilmiştir

Değişkenler

20 Değişken

322 Gözlem

21 KB

AtBat	1986-1987 sezonunda bir beyzbol sopası ile topa yapılan vuruş sayısı
Hits	1986-1987 sezonundaki isabet sayısı
HmRun	1986-1987 sezonundaki en değerli vuruş sayısı
Runs	1986-1987 sezonunda takımına kazandırdığı sayı
RBI	Bir vurucunun vuruş yaptığında koşu yaptırdığı oyuncu sayısı
Walks	Karşı oyuncuya yaptırılan hata sayısı
Years	Oyuncunun major liginde oynama süresi (sene)
CAtBat	Oyuncunun kariyeri boyunca topa vurma sayısı
CHits	Oyuncunun kariyeri boyunca yaptığı isabetli vuruş sayısı
CHmRun	Oyuncunun kariyeri boyunca yaptığı en değerli sayısı
CRuns	Oyuncunun kariyeri boyunca takımına kazandırdığı sayı
CRBI	Oyuncunun kariyeri boyunca koşu yaptırdığı oyuncu sayısı
CWalks	Oyuncunun kariyeri boyunca karşı oyuncuya yaptırdığı hata sayısı
League	Oyuncunun sezon sonuna kadar oynadığı lîgî gösteren A ve N seviyelerine sahip bir faktör
Division	1986 sonunda oyuncunun oynadığı pozisyonu gösteren E ve W seviyelerine sahip bir faktör
PutOuts	Oyun icinde takım arkadaşınla yardımlaşma
Assits	1986-1987 sezonunda oyuncunun yaptığı asist sayısı
Errors	1986-1987 sezonundaki oyuncunun hata sayısı
Salary	Oyuncunun 1986-1987 sezonunda aldığı maaş(bin üzerinden)
NewLeague	1987 sezonunun başında oyuncunun lîgini gösteren A ve N seviyelerine sahip bir faktör

Base Model

```
Call:
lm(formula = Salary ~ ., data = df)

Residuals:
    Min       1Q   Median       3Q      Max
-653.46 -139.03   -5.77  117.33  784.22

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.512e+02  5.034e+02   0.698  0.4861
AtBat       -1.221e+00  1.294e+00  -0.943  0.3466
Hits        6.416e+00  4.983e+00   1.287  0.1993
HmRun       8.223e+00  8.406e+00   0.978  0.3290
Runs       -1.554e+00  3.154e+00  -0.493  0.6227
RBI         -2.929e+00  2.461e+00  -1.190  0.2353
Walks       3.578e+00  1.990e+00   1.798  0.0735 .
Years       3.038e+01  1.310e+01   2.319  0.0213 *
CAtBat      3.360e-02  1.660e-01   0.202  0.8398
CHits      -6.506e-01  7.758e-01  -0.839  0.4026
CHmRun     -3.490e-01  2.012e+00  -0.173  0.8625
CRuns      1.510e+00  9.184e-01   1.644  0.1017
CRBI       8.640e-02  8.880e-01   0.097  0.9226
CWalks     -8.172e-01  3.467e-01  -2.357  0.0193 *
PutOuts    1.164e-01  6.884e-02   1.691  0.0923 .
Assists    1.080e-02  1.786e-01   0.060  0.9518
Errors     -2.135e+00  3.500e+00  -0.610  0.5424
NEW_HitRatio -2.360e+03  1.918e+03  -1.230  0.2198
NEW_RunRatio -2.280e+02  4.016e+02  -0.568  0.5708
NEW_CHitRatio 5.027e+02  2.262e+03   0.222  0.8243
NEW_CRunRatio 3.696e+02  4.691e+02   0.788  0.4316
NEW_Avg_AtBat -1.225e+00  2.226e+00  -0.550  0.5827
NEW_Avg_Hits  8.923e+00  9.272e+00   0.962  0.3369
NEW_Avg_HmRun -4.054e+00  1.796e+01  -0.226  0.8217
NEW_Avg_Runs -8.294e+00  7.600e+00  -1.091  0.2764
NEW_Avg_RBI  3.429e+00  7.658e+00   0.448  0.6547
NEW_Avg_Walks 3.145e+00  3.974e+00   0.792  0.4295
League_N    6.801e+01  6.175e+01   1.101  0.2719
Division_W  -6.544e+01  3.122e+01  -2.096  0.0372 *
NewLeague_N -3.648e+01  6.121e+01  -0.596  0.5519
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 237.9 on 219 degrees of freedom
Multiple R-squared:  0.5263,    Adjusted R-squared:  0.4636
F-statistic:  8.39 on 29 and 219 DF,  p-value: < 2.2e-16
```

Bu model çıktısında "Pr(>|t|)" sütunundaki p değerleri anlamlılığı belirtir.

Çıktıda dikkate alınması gereken anlamlı katsayılar şunlardır:

- Years: p değeri 0.0213
- CWalks: p değeri 0.0193
- Division_W: p değeri 0.0372

1'e yaklaştıkça, modelin veriyi daha iyi açıkladığını gösterir. 0'a yaklaştıkça, modelin veriyi açıklamada başarısız olduğunu gösterir.

Model çıktısındaki R-squared değeri **0.5263**'tür. Bu, modelin verinin **%52.63'ünü** açıkladığı anlamına gelir. Ancak, bu değer yeterince yüksek değildir ve modelin veriyi tam olarak açıklamada sınırlamalara işaret edebilir.

Adjusted R-squared değeri **0.4636**'dır. Bu değer, modeldeki bağımsız değişkenlerin sayısını dikkate alarak R-squared'ı düzeltir. Daha küçük bir değer olması, modeldeki ek değişkenlerin eklenmesinin modelin performansını artırmada sınırlı bir etkisi olduğunu gösterir.

Base Model

```
> vif(model)
```

AtBat	Hits	HmRun	Runs	RBI	Walks	Years
152.134386	203.630193	22.646006	25.697632	16.412444	7.222607	17.137703
CAtBat	CHits	CHmRun	CRuns	CRBI	CWalks	PutOuts
594.012573	1020.085791	102.074133	369.703698	312.781823	34.605782	1.396391
Assists	Errors	NEW_HitRatio	NEW_RunRatio	NEW_CHitRatio	NEW_CRunRatio	NEW_Avg_AtBat
2.915111	2.410398	12.894450	11.130901	11.019910	11.527333	350.341927
NEW_Avg_Hits	NEW_Avg_HmRun	NEW_Avg_Runs	NEW_Avg_RBI	NEW_Avg_Walks	League_N	Division_W
520.805891	55.176808	99.355858	96.766084	17.316845	4.183088	1.069516
NewLeague_N						
4.103166						

Yüksek VIF (10'dan büyük):

AtBat, Hits, CAtBat, CHits, NEW_Avg_AtBat, NEW_Avg_Hits: Bu değişkenlerin VIF değerleri nispeten yüksek, bu da bu değişkenler arasında potansiyel çoklu doğrusallık sorunları olduğunu gösteriyor. Bu, bu değişkenlerin birbirleriyle ilişkili olabileceği anlamına gelir ve bağımlı değişken üzerindeki etkilerini izole etmeyi zorlaştırabilir.

Orta Seviye VIF (5 ile 10 arası):

HmRun, Runs, RBI, Years, CHmRun, CRuns, CRBI, CWalks, PutOuts, Assists, NEW_HitRatio, NEW_RunRatio, NEW_CHitRatio, NEW_CRunRatio, NEW_Avg_HmRun, NEW_Avg_Runs, NEW_Avg_RBI, NEW_Avg_Walks: Bu değişkenlerin VIF değerleri orta seviyede. Tek başlarına sorunlu olmayabilirler, ancak diğer değişkenlerle birlikte etkileşimleri çoklu doğrusallığa neden olabilir.

Düşük VIF (5'ten küçük):

Walks, Errors, League_N, Division_W, NewLeague_N: Bu değişkenlerin VIF değerleri nispeten düşük, bu da onların çoklu doğrusallıktan daha az etkilendiğini gösteriyor.

Değişken Seçim Yöntemleri

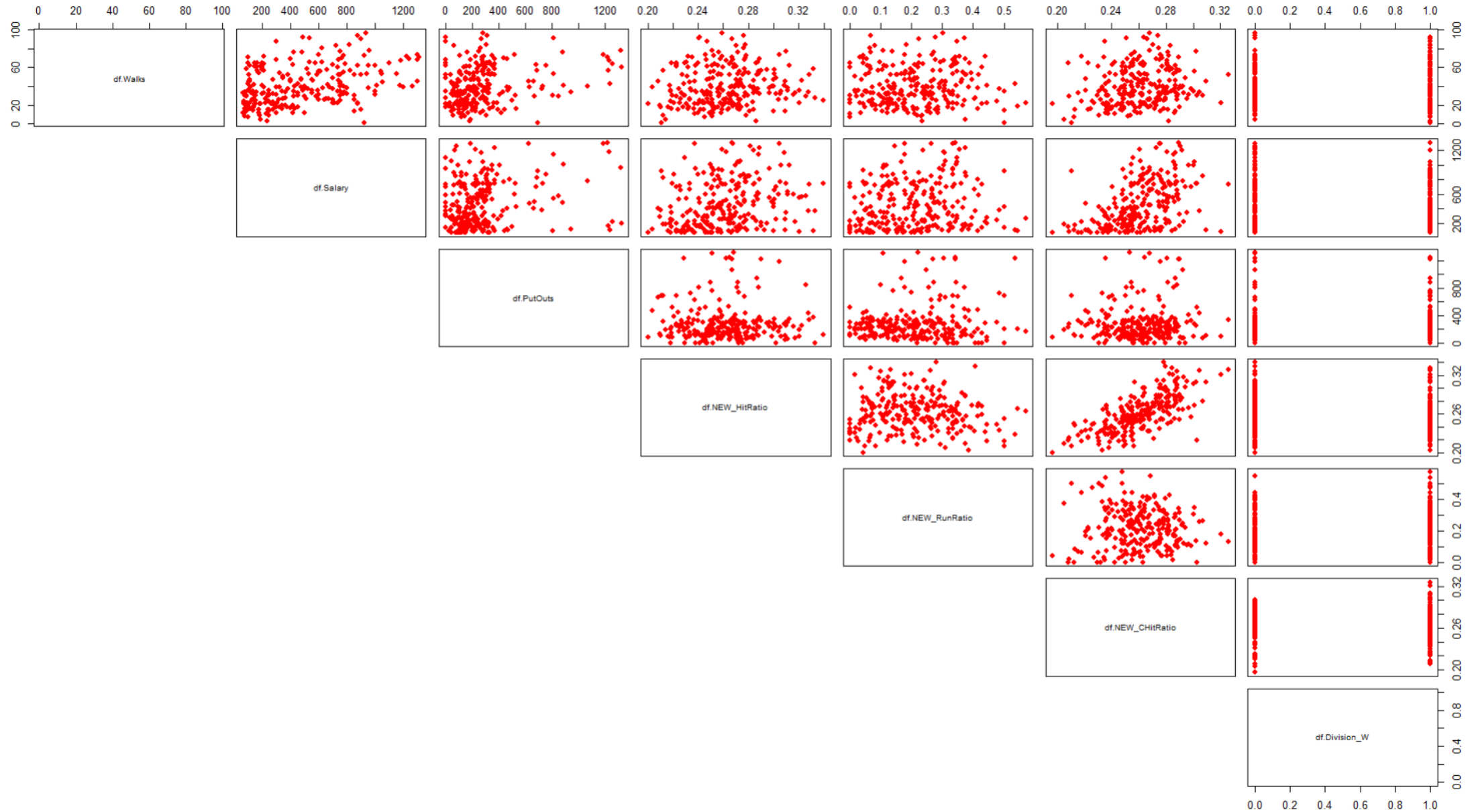
	mindex	n	predictors	rsquare	adjr	predrsq	cp	aic	sbic	sbc	msep
1117	1024	6	Walks PutOuts NEW_HitRatio NEW_RunRatio NEW_CHitRati...	0.3598549	0.3439835	0.3156312	4.604920	3490.552	2784.467	3518.692	17230662
739	562	5	Walks NEW_HitRatio NEW_RunRatio NEW_CHitRatio Divisio...	0.3529511	0.3396373	0.3194339	5.189001	3491.223	2784.928	3515.846	17344521
1523	1486	7	Walks PutOuts Assists NEW_HitRatio NEW_RunRatio NEW_C...	0.3621361	0.3436089	0.3130079	5.751084	3491.664	2785.715	3523.321	17240798

Değişken Seçim Yöntemleri

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	NEW_CHitRatio	addition	0.218	0.214	50.7640	3530.5081	287.8442
2	walks	addition	0.313	0.308	16.5620	3499.9967	270.2003
3	NEW_RunRatio	addition	0.330	0.321	12.4030	3496.0169	267.5196
4	NEW_HitRatio	addition	0.344	0.333	8.9940	3492.6445	265.1909

Değişken Seçim Yöntemleri



Final Model

```
Call:
lm(formula = Salary ~ Walks + PutOuts + NEW_HitRatio + NEW_RunRatio +
    NEW_CHitRatio + Division_W, data = df)

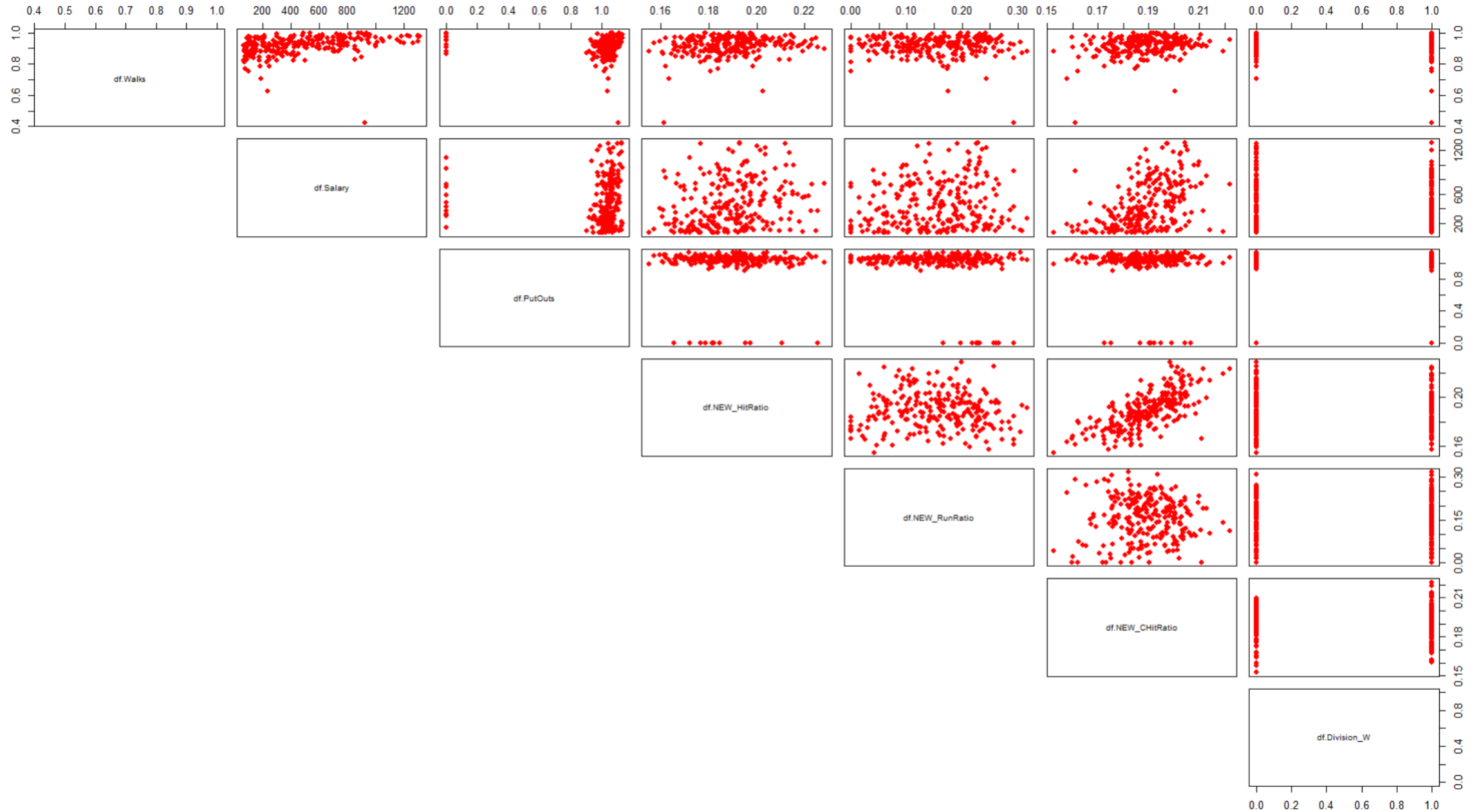
Residuals:
    Min       1Q   Median       3Q      Max
-595.1 -193.7  -16.2  159.4 1050.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1835.86    247.05  -7.431 1.85e-12 ***
Walks         139.51     30.26   4.610 6.52e-06 ***
PutOuts        18.16     12.93   1.404 0.16148
NEW_HitRatio -2516.29   1079.92  -2.330 0.02063 *
NEW_RunRatio   458.02    168.65   2.716 0.00709 **
NEW_CHitRatio 9646.59   1391.17   6.934 3.70e-11 ***
Division_W     -57.86     34.35  -1.685 0.09334 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

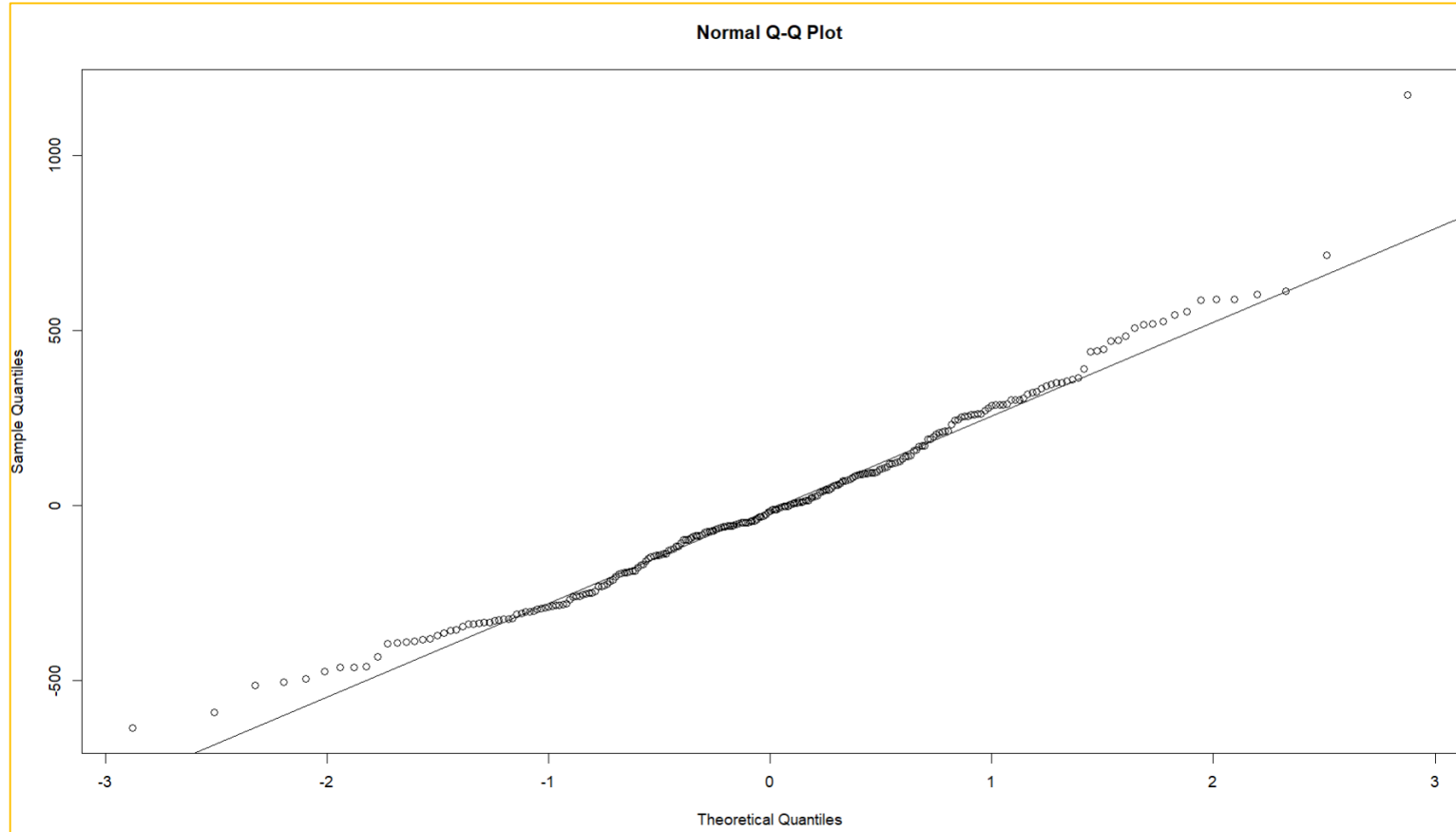
Residual standard error: 269.5 on 242 degrees of freedom
Multiple R-squared:  0.3282,    Adjusted R-squared:  0.3115
F-statistic: 19.7 on 6 and 242 DF,  p-value: < 2.2e-16
```

```
> vif(log_transform_model)
```

Walks	PutOuts	NEW_HitRatio	NEW_RunRatio	NEW_CHitRatio	Division_W
1.083714	1.052867	1.988707	1.053812	2.062431	1.009091



Hataların Normal Dağıldığı Varsayımı: Q-Q Plot



```
> ks.test(log_transform_model$residuals, "pnorm")
```

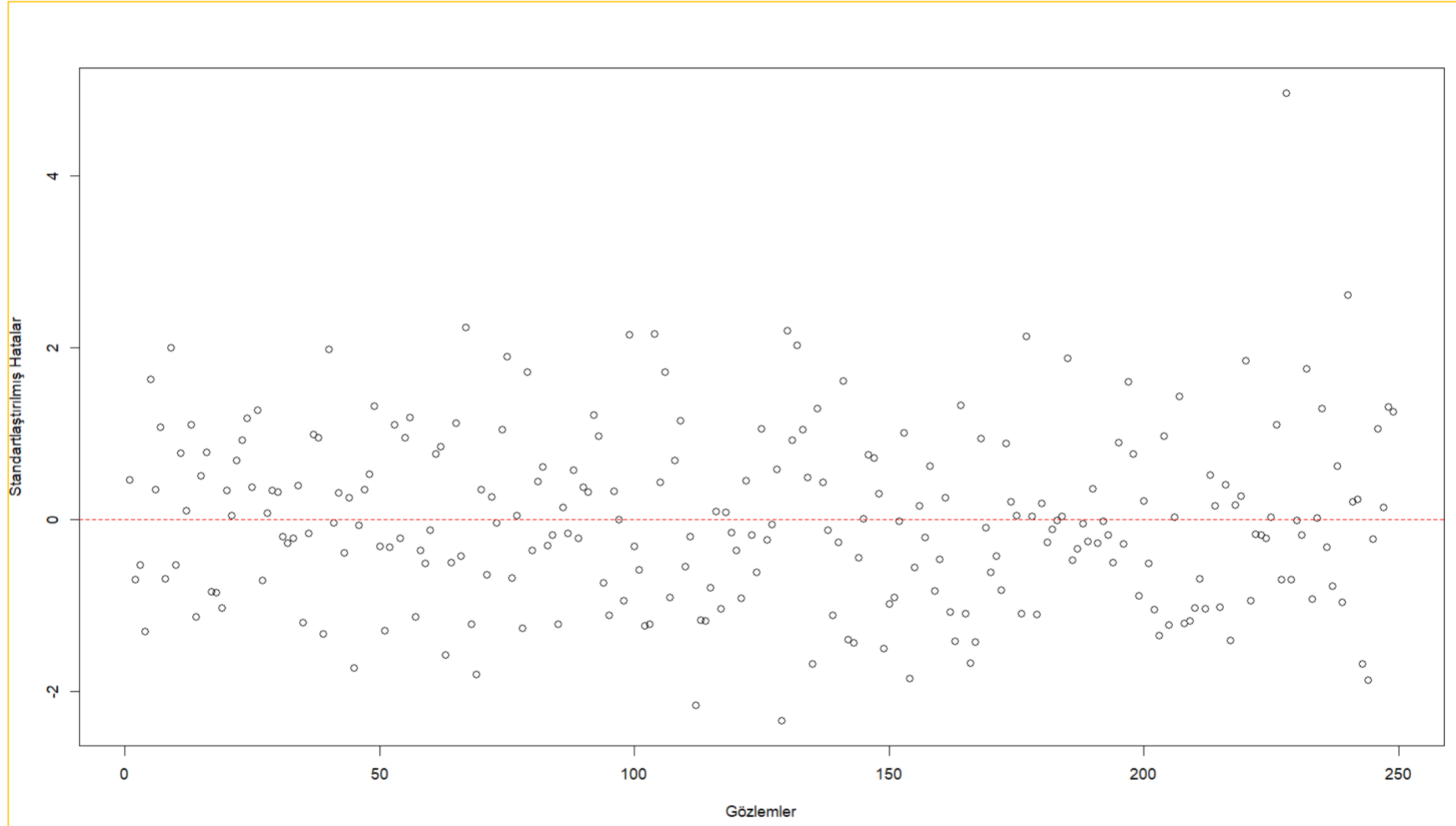
Asymptotic one-sample Kolmogorov-Smirnov test

data: log_transform_model\$residuals

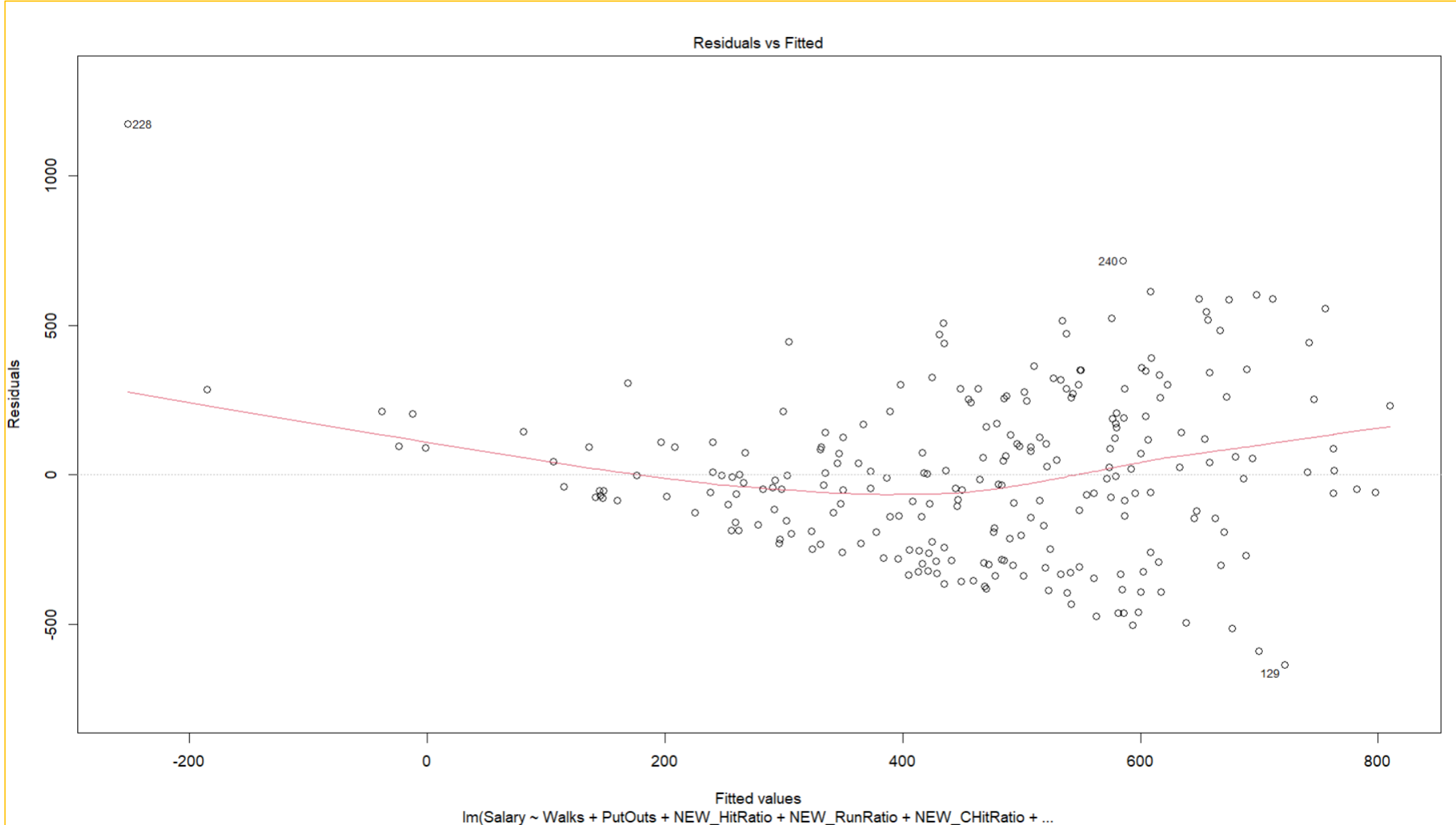
D = 0.52566, p-value < 2.2e-16

alternative hypothesis: two-sided

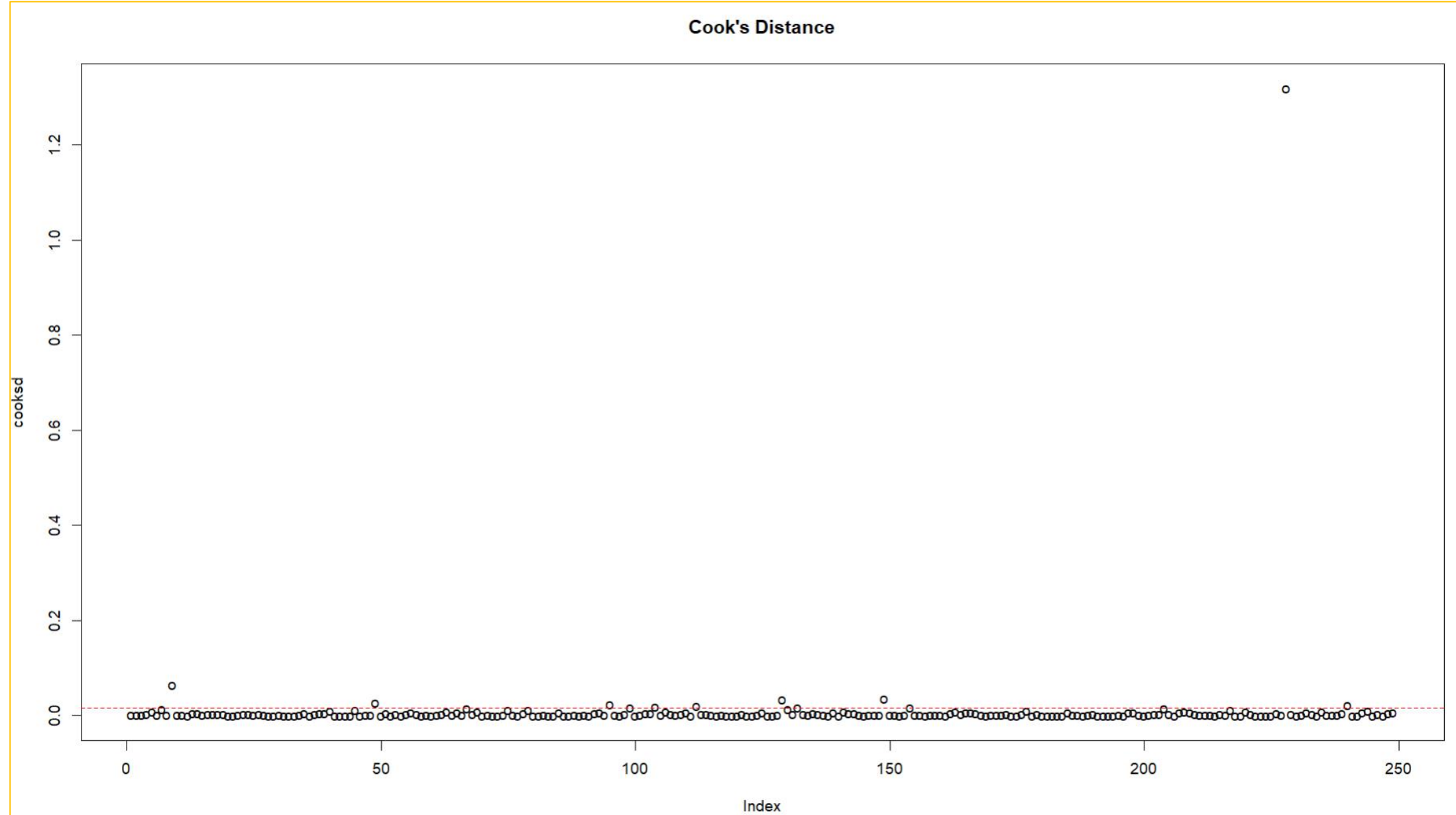
Hataların Sabit Varyanslı Hali: Grafik ve Test Sonuçları



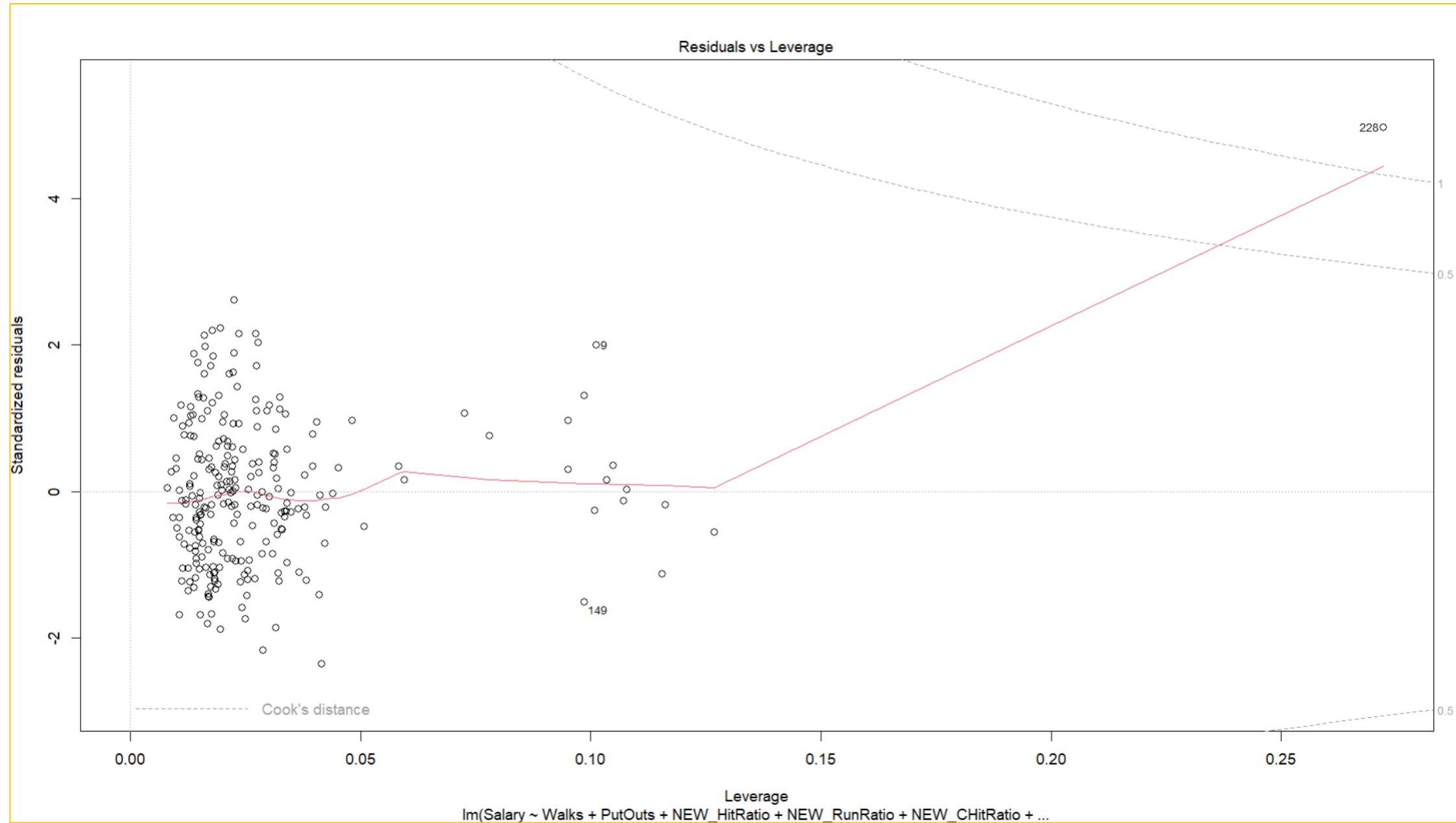
Hataların Sabit Varyanslı Hali: Artık Grafiği



Cook's Distance



Residuals vs. Leverage



Güven Aralıkları

```
> confint(log_transform_model)
              2.5 %      97.5 %
(Intercept) -3310.5778 -1876.73659
Walks        383.2049  1545.42118
PutOuts      -113.4489   210.33087
NEW_HitRatio -7071.7423 -548.00078
NEW_RunRatio  164.9461  1117.21691
NEW_CHitRatio 10481.4175 18850.47500
Division_W    -125.9986   12.87233
> confint(log_transform_model, level=0.95)
              2.5 %      97.5 %
(Intercept) -3310.5778 -1876.73659
Walks        383.2049  1545.42118
PutOuts      -113.4489   210.33087
NEW_HitRatio -7071.7423 -548.00078
NEW_RunRatio  164.9461  1117.21691
NEW_CHitRatio 10481.4175 18850.47500
Division_W    -125.9986   12.87233
```

%95 Güven Aralığında Yeni Gözlem Tahmin Sonuçları

```
> # Yeni gözlem değeri için 95% güven aralığı
> new_observation <- data.frame(walks = 0.99,
+                               PutOuts = 1.11,
+                               NEW_HitRatio = 0.20,
+                               NEW_RunRatio = 0.29,
+                               NEW_CHitRatio = 0.20,
+                               Division_W = 1)
> predict_interval <- predict(log_transform_model, newdata = new_observation, interval = "confidence", level = 0.95)
> predict_interval
      fit      lwr      upr
1 715.3476 615.5252 815.1701
```

95% Güven Aralığı: 715.3476 ile 615.5252

95% Kestirim Aralığı: 715.3476 ile 161.6556