Methoden der Schließenden Statistik 2

Ranjit Sah (836261)
Rahul Kandel (827352)

Hausaufgabenblatt SoSe 2019

Einleitung:

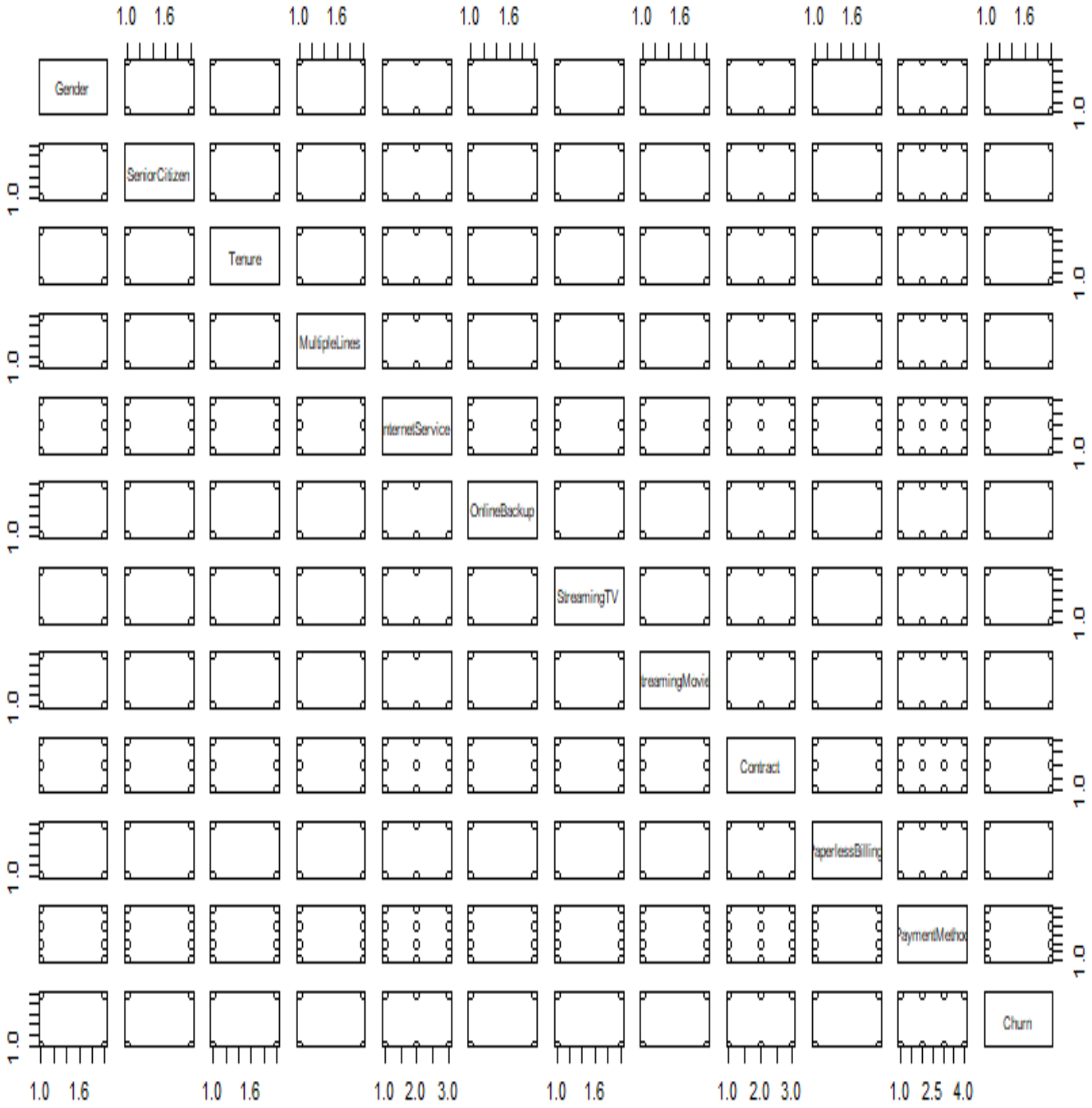
im dem folgenden Bericht wir machen Datenauswertung durch logistische Regression. Wir analysieren Der Datensatz Telecom mit einem logistischen Regressionsmodell und danach unterschiedliche Methoden überprüft, um zu sehen welches Modell zum Datensatz besser passt. Eine logistische Regression kann in R mit der Funktion glm () gerechnet werden. Wichtig dabei ist, dass als Familie binomial angegeben wird. Doch vor dem rechnen eine Regression muss zuerst der Datensatz eingelesen und rekodiert werden.

variable des Datensatzes:

Es wird der Datensatz Telekom gelesen. Der Datensatz enthält 1000 Zeilen und 12 Variablen. Jede Zeile bezieht sich auf einem Kunden (Datenelement). Alle Variablen sind Gruppiert (Faktorvariablen). Die Zielgröße ist Churn. „Yes“= der Kunde ist abgewandert, „No“= der Kunde bleibt.

```
> head(Telecom)#um einen kurzen Blick auf die Daten zu werfen
     Gender SeniorCitizen Tenure MultipleLines InternetService OnlineBackup
2533 Female           Yes   TRUE           No     Fiber optic          No
3113   Male            No   TRUE          Yes     Fiber optic         Yes
6482 Female            No  FALSE           No     Fiber optic          No
5508 Female            No   TRUE           No             DSL         Yes
2544   Male            No   TRUE          Yes     Fiber optic          No
1292   Male            No  FALSE           No     Fiber optic          No
     StreamingTV StreamingMovies       Contract PaperlessBilling
2533         Yes             Yes       One year              Yes
3113          No             Yes       Two year              Yes
6482          No              No Month-to-month               No
5508          No              No       One year              Yes
2544         Yes             Yes Month-to-month               No
1292          No              No Month-to-month               No
                PaymentMethod Churn
2533 Bank transfer (automatic)    No
3113 Bank transfer (automatic)    No
6482   Credit card (automatic)    No
5508          Electronic check    No
2544              Mailed check   Yes
1292          Electronic check   Yes
```

Statistiken:



```
> summary(Telecom) #summary  für all variable
    Gender    SeniorCitizen   Tenure    MultipleLines    InternetService OnlineBackup
 Female:479   No :828       FALSE:303   No :572       DSL         :367    No :632
 Male  :521   Yes:172       TRUE :697   Yes:428       Fiber optic:437    Yes:368
                                                      No          :196

 StreamingTV StreamingMovies           Contract   PaperlessBilling
 No :592     No :607         Month-to-month:558   No :402
 Yes:408     Yes:393         One year      :212   Yes:598
                             Two year      :230

                  PaymentMethod Churn
 Bank transfer (automatic):246   No :741
 Credit card (automatic)  :200   Yes:259
 Electronic check         :329
 Mailed check             :225
> prop.table(table(Telecom$Churn))#checking out the percentages

   No   Yes
0.741 0.259
```

Modellentwicklung durch AIC:

Nach einer explorativen Analyse der Daten und der Wahl einer passenden Modellklasse, geht es darum, das bestmögliche Modell zu den vorliegenden Daten zu finden. Daher stellt sich die Frage, was "bestmögliches" Modell bedeutet und wie ein solches bestimmt werden kann.

Das Akaike-Informationskriterium (AIC) (Akaike, 1974) ist eine Geldbuße-Technik, die auf der Anpassung in der Stichprobe basiert, um die Wahrscheinlichkeit eines Modells zu schätzen, die zukünftigen Werte vorherzusagen / abzuschätzen. Ein gutes Modell hat unter allen anderen Modellen einen minimalen AIC.

Ein Modell, das durch Devianz Tests entwickelt wird, kann zu Überanpassung führen. Entweder eine Variable im Modell ist statistisch signifikant aber die Größe des Effekts ist geringfügig, oder das Modell ist nicht reproduzierbar — Wenn die Datenerhebung wiederholt würde, würde man ein anderes Modell bekommen.

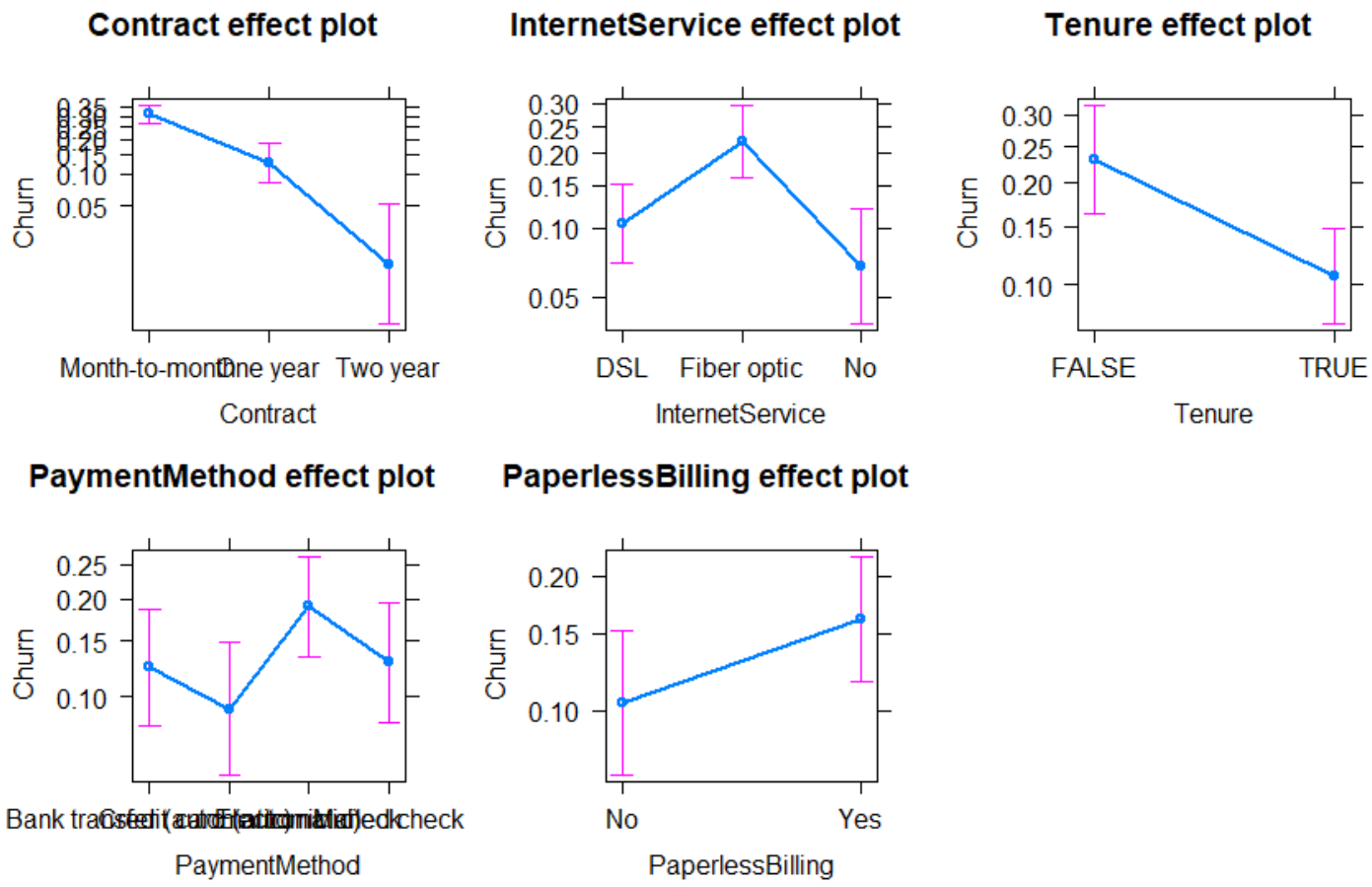
Endmodell:

Wenn der AIC-Verbesserung klein ist, ist manchmal der Devianz Test nicht signifikant. Um eine Variable in dem Modell einzubringen, sollte den AIC-Wert kleiner und die Devianz signifikant zu einem gegebenen α-Niveau sein. Man findet das beste Haupteffektmodell, bevor man Wechselwirkungen untersucht. Mit Hilfe Vorlesung Skript 8 Am Ende wir haben End Model mit zielgrösse churn und Einflussgröße Contract, InternetService, Tenure,

Payment-method and PaperlessBilling gebaut.d.h Aktuall_mod5.

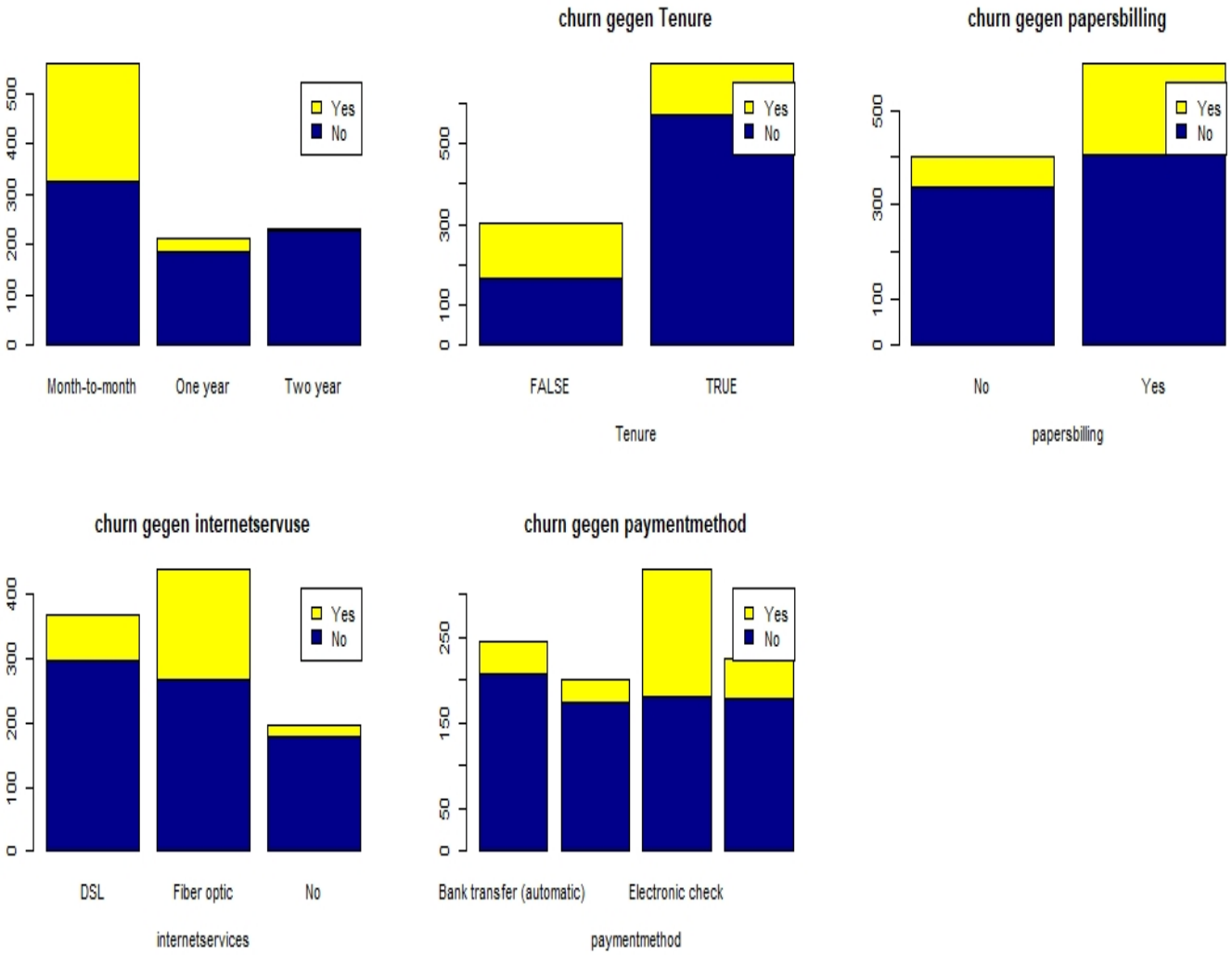
```
> Aktuall_mod5<-glm(Churn~Contract+InternetService+Tenure+PaymentMethod+PaperlessBilling,data=Telecom,family = binomial())
> summary(Aktuall_mod5)

Call:
glm(formula = Churn ~ Contract + InternetService + Tenure + PaymentMethod +
    PaperlessBilling, family = binomial(), data = Telecom)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6429  -0.7160  -0.2621   0.7749   3.3981

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -0.86067    0.28213  -3.051  0.00228 **
ContractOne year                 -1.13854    0.25759  -4.420 9.87e-06 ***
ContractTwo year                 -3.51155    0.72653  -4.833 1.34e-06 ***
InternetServiceFiber optic        0.89447    0.20100   4.450 8.58e-06 ***
InternetServiceNo                -0.46792    0.31221  -1.499  0.13394
TenureTRUE                       -0.93030    0.19955  -4.662 3.13e-06 ***
PaymentMethodCredit card (automatic) -0.34798    0.30622  -1.136  0.25580
PaymentMethodElectronic check     0.50554    0.23698   2.133  0.03290 *
PaymentMethodMailed check         0.04355    0.28970   0.150  0.88051
PaperlessBillingYes               0.50996    0.19202   2.656  0.00791 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1144.02  on 999  degrees of freedom
Residual deviance:  846.23  on 990  degrees of freedom
AIC: 866.23

Number of Fisher Scoring iterations: 7
```

Zusammenfassung aller Einflussgröße im Endmodell(Grafik):

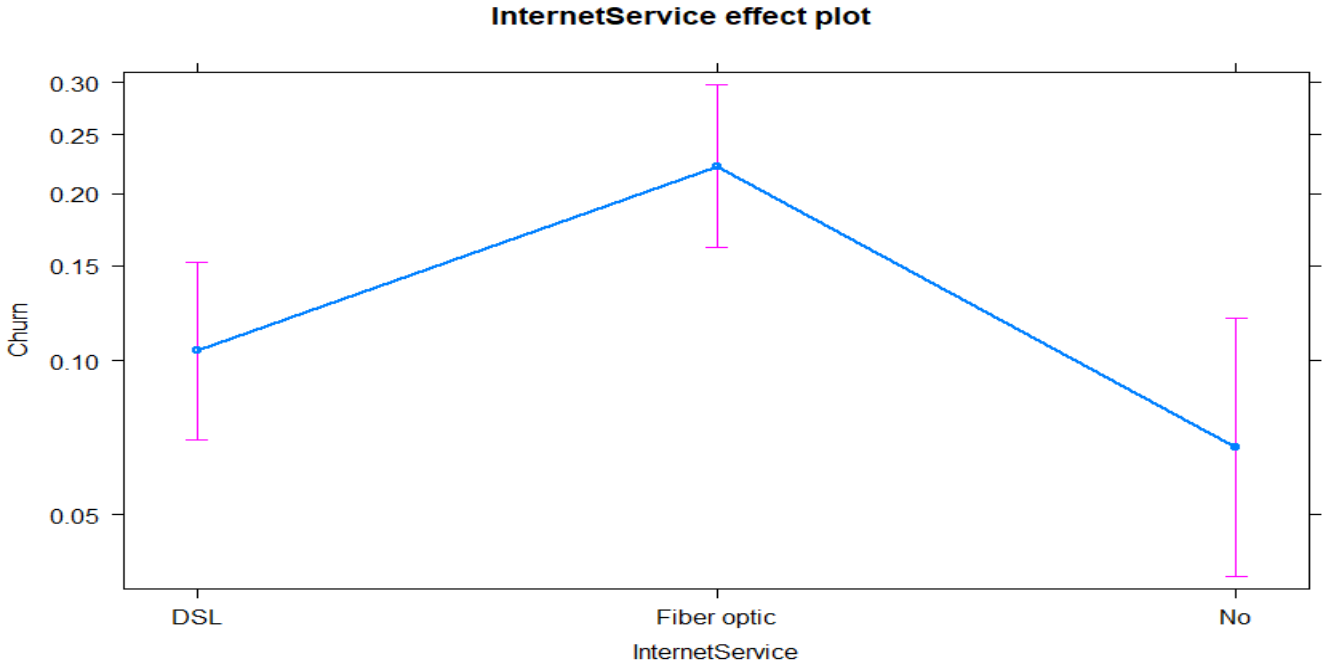


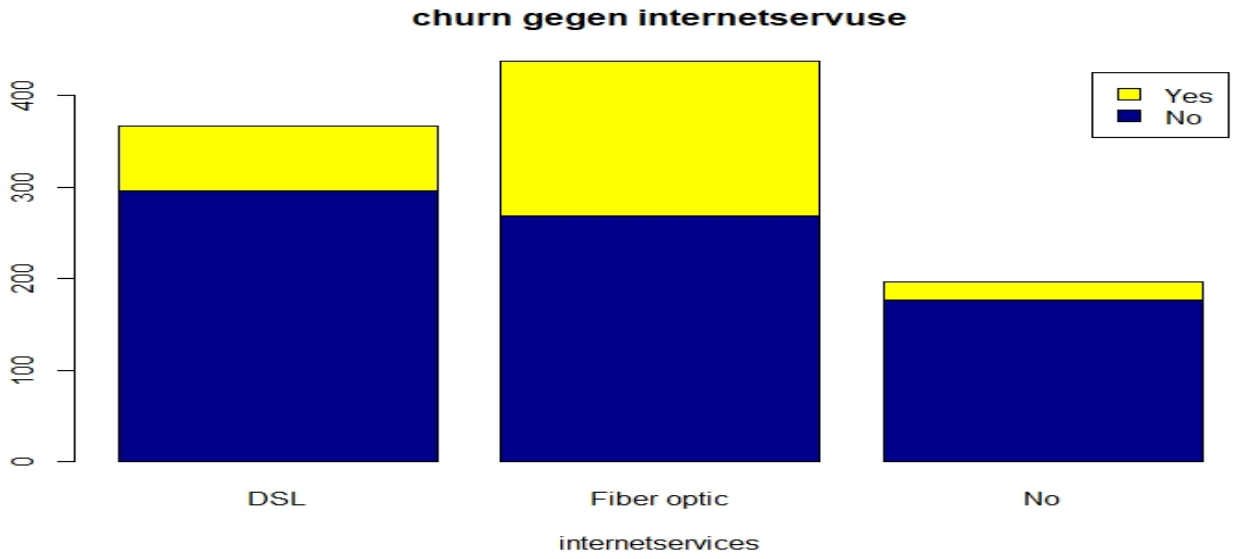
in der obigen Bild ist ein Main Effet Plot von end Modelle gibt 5 faktorvariable signifikant. Dass ist Contract:im diese Bild wir sehen Month-to-month contract hat maximum 0.30 probabilty,one year hat 0.14 und Two Year hat weniger als 0.005 gegen churn.Tenure hat 0.22 False und 0.12 True probability gegen churn.Payment-method :0.12 Prozent bezahlt bei Bank Transfer, bezahlt bei credit Card weniger als 0.9,0.19 bei Elektronik check und weniger als 0.15 bei Mailed check.PaperlessBilling:0.16 Prozent enthalt Papers Billing und 0.11 hat kein Papers Billing.

Boxplot End Modell:



Internet Service: im diese Bild wir sehen DSL hat 0.11 probabilty,Fiber Optik hat 0.45 und 0.07 hat keine Dsl und Fiber Optik gegen churn





Interpretation:

Eine Möglichkeit der Interpretation von logistischen Modellen ist das berechnen von vorausgesagten Wahrscheinlichkeiten und diskreten Änderungen. Das geht ganz einfach mit dem Paket glm. predict

```
library(glm.predict)
predicts(Aktuall_mod5,"0,1;F", position = 1)
val1_mean val1_lower val1_upper val2_mean val2_lower val2_upper dc_mean dc_lower dc_upper Contract_val1 Contract_val2
      NaN         NA         NA       NaN         NA         NA     NaN       NA       NA            NA            NA
      NaN         NA         NA       NaN         NA         NA     NaN       NA       NA            NA            NA
      NaN         NA         NA       NaN         NA         NA     NaN       NA       NA            NA            NA
InternetService Tenure PaymentMethod PaperlessBilling
             NA      0             1              DSL
             NA      0             1      Fiber optic
             NA      0             1               No
```

ODDS Ratio: Bei logistischen Modellen kommtzusätzlich die Interpretationsmöglichkeit der Odds Ratio (Quotenverhältnis) hinzu.

```
> #odd Ratio fur End Modell
> OR<-exp(coef(Aktuall_mod5))
> OR
                          (Intercept)                     ContractOne year                     ContractTwo year
                            0.42287662                          0.32028630                          0.02985048
           InternetServiceFiber optic                     InternetServiceNo                          TenureTRUE
                            2.44603064                          0.62630317                          0.39443568
PaymentMethodCredit card (automatic)         PaymentMethodElectronic check             PaymentMethodMailed check
                            0.70611168                          1.65788615                          1.04451209
                  PaperlessBillingYes
                            1.66521650
>
```

LOGoddRatio:

```
> log(OR)
                          (Intercept)                     ContractOne year                     ContractTwo year
                           -0.86067481                         -1.13853999                         -3.51155438
           InternetServiceFiber optic                     InternetServiceNo                          TenureTRUE
                            0.89446656                         -0.46792073                         -0.93029918
PaymentMethodCredit card (automatic)         PaymentMethodElectronic check             PaymentMethodMailed check
                           -0.34798187                          0.50554339                          0.04354988
                  PaperlessBillingYes
                            0.50995514
```

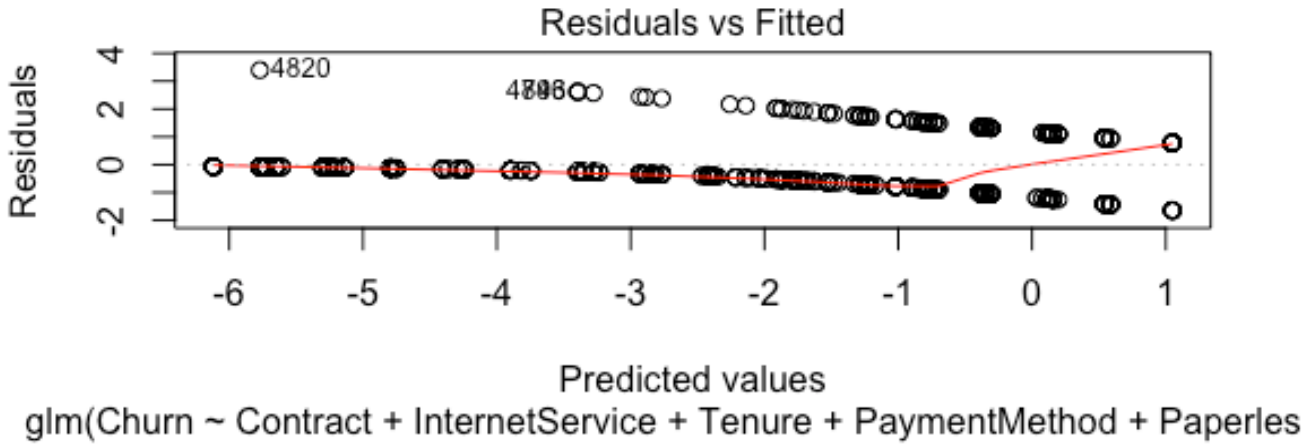
Anova Modell:

Ein Anova Modell ist eine Art linearen Modells, dessen Einflussvariablen nominal oder ordinal skaliert sind. Wir haben bereits zwei Anova-Modelle gelernt, einfaktorielle und zweifaktorielle Anova. In einfaktorieller Anova hängt die Zielgröße von einer Variablen ab und in zweifaktorieller Anova (auch Haupteffekktmodell benannt) hängt die Zielgröße von zwei Variablen ab. Und es gibt auch die zweifaktorieller Anova mit Wechselwirkung, die von beiden Variablen und der Kombination von ihnen abhängt. In unserem Beispiel haben wir ein zweifaktorielles Anova Modell und deshalb muss man in dem Fall zuerst die Wechselwirkung als Modell prüfen, die wir auch gemacht haben und da es nicht signifikant war, haben wir es weggelassen. Das heißt, wir haben jetzt ein Haupeffektmodell

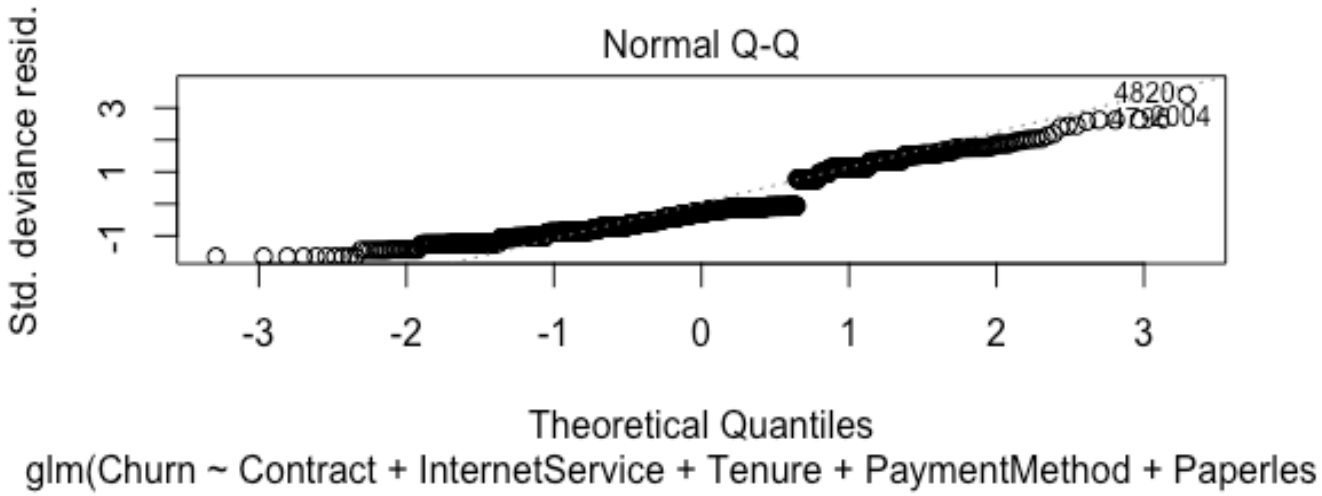
```
> anova(Aktuall_mod5,test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                           999    1144.02
Contract        2  212.996       997     931.02 < 2.2e-16 ***
InternetService 2   39.510       995     891.51 2.633e-09 ***
Tenure          1   24.369       994     867.14 7.954e-07 ***
PaymentMethod   3   13.728       991     853.41  0.003299 **
PaperlessBilling 1    7.182       990     846.23  0.007363 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Diagnostiken Diagramme:

- Zum Schluss werden wir die folgenden Eigenschaften überprüfen:
- Residuen
- Heteroskedastizität
- Residuen vs. Leverage

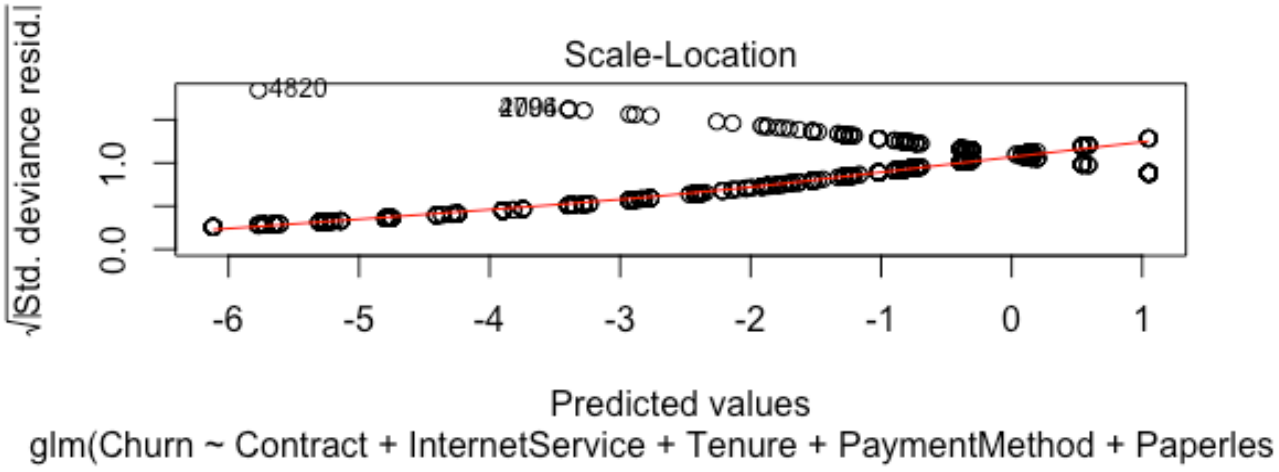


Man sieht eine gerade Linie und die Werte liegen mehr oder weniger nahe an der Linie, also es ist angemessen.



Das Residuen-QQ-Diagramm ist auch sehr ordentlich und Werte außerhalb -3 und + 3 sieht man nicht.

2-Heteroskedastizität:



In dem Scale-Location-Diagramm ist eine gerade Linie gut zu sehen und das ist ein gutes Kriterium, was unser Modell anbelangt.

3-Residuen vs. Leverag: