Lennart Oelschläger (lennart.oelschlaeger@uni-bielefeld.de)
310214 Praktische Übung zu "Einführung in die Mikroökonometrie"
Wintersemester 2024/2025

Aufgabenblatt 5

Lösungen

Überblick

In einem *binären Wahlmodell* ist die abhängige Variable binär. Wir modellieren sie nicht direkt, sondern stattdessen die Wahrscheinlichkeit, dass sie den Wert 1 annimmt (dass sie den Wert 0 annimmt, folgt dann direkt aus der Gegenwahrscheinlichkeit). Lineare Modelle weisen hier Nachteile auf: Prognosen können negativ sein oder oberhalb von 1 liegen und es liegt zwangsläufig Heteroskedastizität vor. Als bessere Alternative gibt es *Logit-* und *Probit-Modelle*, sie werden mit der Maximum-Likelihood Methode geschätzt. Mit *ROC* Kurven können die Klassifizierungsgüten verschiedener Modelle verglichen werden.

Fortsetzung Aufgabenblatt 4

1. Angenommen, die abhängige Variable y ist nun binär. Welche der Annahmen im klassischen linearen Regressionsmodells gelten dann nicht mehr?

- MLR.1-4 können weiterhin erfüllt sein
- MLR.5 und MLR.6 gelten nicht mehr (siehe Vorlesung)

2. Welche neue Interpretation bekäme $\hat{\beta}_1$?

Verändert sich x, so verändert sich nicht der (relative) Wert von y, sondern die **Wahrscheinlichkeit**, dass $y = 1$ eintritt. Jedoch sind die Änderungen nicht mehr linear, sodass eine gewohnte ceteris paribus nicht möglich ist und wir andere Methoden (APE, PEA, etc.) benötigen.

Aufgaben

1. Simulieren Sie Daten wie nachfolgend angegeben und schätzen Sie die Probit-Modelle

$$y = \beta_0 + \beta_1 x_1 + u \tag{1}$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \tag{2}$$

   mit der Funktion `glm()`.

```
set.seed(1)
data <- MASS::mvrnorm(n = 50, mu = c(0, 0, 0), Sigma = diag(3))
data <- as.data.frame(data)
colnames(data) <- c("x_1", "x_2", "u")
```

```
beta_0 <- 1
beta_1 <- 0.2
beta_2 <- -0.5
y <- beta_0 + data$x_1 * beta_1 + data$x_2 * beta_2 + data$u
data$y <- ifelse(y >= 0.5, 1, 0)
```

```
R> mod1 <- glm(y ~ x_1, data = data, family = binomial(link = "probit"))
R> mod2 <- glm(y ~ x_1 + x_2, data = data, family = binomial(link = "probit"))
```

2. Welches der Modelle (1) und (2) hat die “bessere” Null Deviance beziehungsweise Residual Deviance und was folgt daraus?

```
R> data.frame(
+   "null.deviance" = c(mod1$null.deviance, mod2$null.deviance),
+   "residual.deviance" = c(mod1$deviance, mod2$deviance),
+   row.names = c("model 1", "model 2")
+ )
```

```
#         null.deviance residual.deviance
# model 1      62.68695          62.65869
# model 2      62.68695          52.59964
```

Je kleiner die Deviance, desto besser beschreibt das Modell die Daten. Die Null Deviance der Modelle ist identisch, denn ohne Regressoren sind beide Modelle identisch. Die Residual Deviance von Modell (2) ist kleiner als die von Modell (1), also erklärt Modell (2) die Variable y besser als das Modell (1).

3. Bitte testen Sie für beide Modelle die Hypothese $H_0 \colon \beta_k = 0, k = 1, \ldots, K$ zu $\alpha = 10\%$.

Diese Hypothese lässt sich mit der Teststatistik "Null Deviance minus Residual Deviance" testen, die unter der Nullhypothese Chi-Quadrat verteilt ist mit K Freiheitsgraden. Sofern der Wert der Teststatistik größer ist als das $1 - \alpha = 90\%$ Quantil der Chi-Quadrat Verteilung mit K Freiheitsgraden, wird die Nullhypothese verworfen.

```
R> # Modell 1: H_0 kann nicht verworfen werden
R> mod1$null.deviance - mod1$deviance > qchisq(p = 0.9, df = 1)
```

```
# [1] FALSE
```

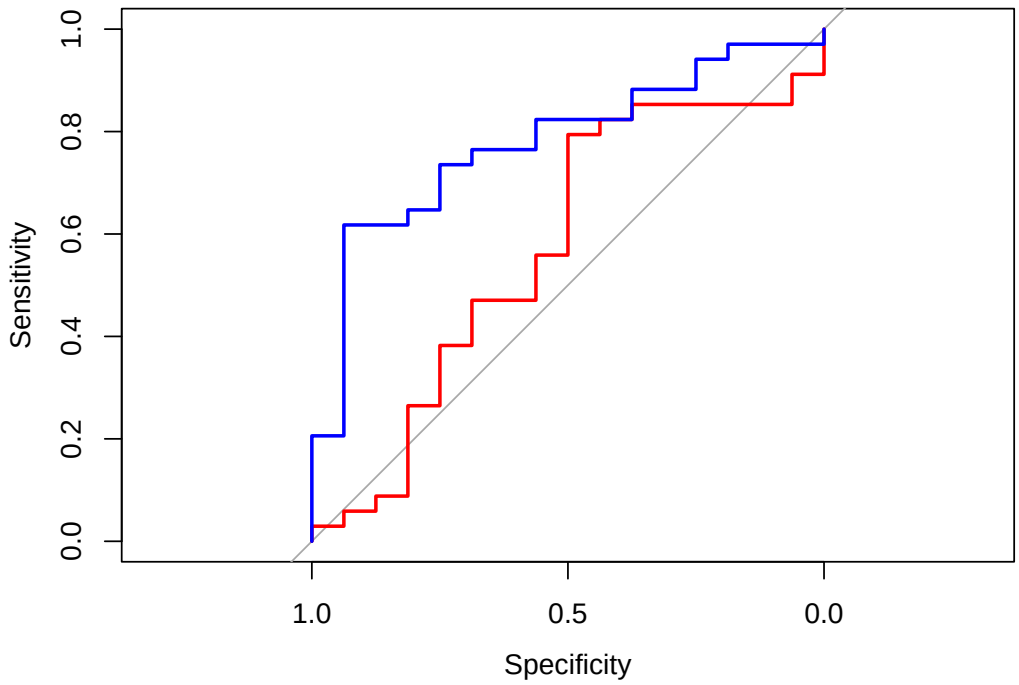
```
R> # Modell 2: Verwerfe H_0
R> mod2$null.deviance - mod2$deviance > qchisq(p = 0.9, df = 2)
```

```
# [1] TRUE
```

4. Zeichen Sie für beide Modelle die ROC Kurve mit der Funktion `pROC::roc()`. Welches Modell hat demnach die bessere Prognosequalität?

```
R> pROC::roc(
+   data$y, pnorm(predict(mod1)), plot = "TRUE", col = "red", quiet = TRUE
+ )
```

```
R> pROC::roc(
+   data$y, pnorm(predict(mod2)), plot = "TRUE", col = "blue", quiet = TRUE,
+   add = TRUE
+ )
```



Die blaue Kurve liegt oberhalb der roten Kurve, daher hat das Modell 2 durchgängig die bessere Sensitivität und Spezifität.