Exam 4

Johan Bysted, Jonathan Arve og Mathias Kold

2024-05-30

Exam 4 - Models for binary variables

In a multiple linear regression (MLR) there are 6 assumptions. Assumption 1-5 are called Gauss-Markov assumptions and assumption 6 is called the normality assumption. The first 4 assumptions exist to secure that the model is unbiased, while assumption 5 checks for heteroskedasticity and assumption 6 checks for normality in the model. The assumptions are:

MLR1: Linear in Parameters

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the unknown parameters of interest and u is an unobserved random error or disturbance term.

MLR2: Random Sampling

We have a random sample of n observations, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, 2, \ldots, n\}$, following the population model in Assumption MLR.1.

MLR3: No Perfect Collinearity

In the sample (and therefore in the population), none of the independent variables is constant, and there are no exact linear relationships among the independent variables.

MLR4: Zero Conditional Mean

The error u has an expected value of zero given any values of the independent variables. In other words,

$$E(u|x_1, x_2, \ldots, x_k) = 0$$

MLR5: Homoskedasticity

The error u has the same variance given any value of the explanatory variables. In other words,

$$\text{Var}(u|x_1, x_2, \ldots, x_k) = \sigma^2$$

MLR6: Normality

The population error u is independent of the explanatory variables $x_1, x_2, \ldots, x_k$ and is normally distributed with zero mean and variance σ^2: $u \sim \text{Normal}(0, \sigma^2)$.

In this assignment, we investigate the factors that influence whether women in Switzerland participate in the labor force.

The dependent variable is participation, a binary variable that measures whether the person is part of the labor force. Additionally, we have seven explanatory variables: income that is not work-related measured in 1000 CHF (income), age (age), age squared (agesq), education measured in years (educ), number of children under 7 years old (youngkids), number of children over 7 years old (oldkids), and a dummy variable indicating whether the person is a foreigner (foreign).

The dataset data4 contains these variables measured for 872 Swiss women.

1. Set up a linear regression model for participation where you use the described explanatory variables.

(a) Estimate the model using OLS and comment on the results.

Here we use the lm() function to estimate the model.

```
model <- lm(part~inc+age+age_sq+educ+you+old+fore)
summary(model)
```

```
##
## Call:
## lm(formula = part ~ inc + age + age_sq + educ + you + old + fore)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84324 -0.39866 -0.08992  0.42048  1.01049
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.3685651  0.2529630  -1.457  0.14548
## inc         -0.0035163  0.0007098  -4.954 8.75e-07 ***
## age          0.0633852  0.0128603   4.929 9.92e-07 ***
## age_sq      -0.0009029  0.0001566  -5.767 1.12e-08 ***
## educ         0.0067725  0.0059615   1.136  0.25626
## you         -0.2390033  0.0313780  -7.617 6.82e-14 ***
## old         -0.0474930  0.0171593  -2.768  0.00576 **
## fore         0.2572106  0.0401252   6.410 2.39e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4506 on 864 degrees of freedom
## Multiple R-squared:  0.1901, Adjusted R-squared:  0.1836
## F-statistic: 28.98 on 7 and 864 DF,  p-value: < 2.2e-16
```

From the model, it can be seen that the income variable has an estimate of -0.0035 which means that one unit (1000 CHF) more income will give approximately 0.35% lower probability of participation.

The age variable has an estimate of 0.0634 while the age squared variable has an estimate of -0.0009 which means that the probability of participation will be higher up to a certain age where the probability will start to drop. This makes sense due to people retiring at a certain age.

The education variable has an estimate of 0.0068 which means that one extra year of education will give approximately a 0.68% higher probability of participation.

The young kids variable for number of children under 7 years old has an estimate of -0.239 which means that if you have children under 7 years old, you will have approximately 23.9% lower probability of participation than people without children under 7 years old.

The old kids variable for number of children over 7 years old has an estimate of -0.0475 which means that if you have children over 7 years old, you will have approximately 4.75% lower probability of participation than people without children over 7 years old.

The foreign variable has an estimate of 0.2572 which means that you have approximately 25.72% higher probability of participation if you are foreign than if you are not foreign.

The intercept is -0.0035 which is the value of the dependent variable, in this case participation in the labor force, when all other variables have a value of 0.

It can also be seen that all the variables except education are statistically significant at a 5% significance level due to p-values < 0.05. All variables except education are also significant at a 1% significance level.

(b) Test whether the partial effect of education is different from zero.

Using a t-test, we can test whether the partial effect of education is different from zero. We set up our null hypothesis:

$$H_0 : \beta_4 = 0$$

And the alternative hypothesis:

$$H_1 : \beta_4 \neq 0$$

If the t-value is within the 95% accept-interval, we fail to reject our null hypothesis. Otherwise, we reject our null hypothesis, and instead accept our alternative hypothesis. We perform a two-sided test, which means there is 2.5% outside of the acceptance interval on each side.

The t-score can be calculated as follows:

$$t = \frac{\beta_j}{se(\beta_j)}$$

The coefficient, β is given from the summary of the model. The same goes for the standard error.

```
beta <-  0.0067725; se <- 0.0059615
tscore <- beta/se
tscore
```

```
## [1] 1.13604
```

To calculate the accept interval, we use the qt() function and our degrees of freedom

```
df <- length(data$educ)-1
acc <- qt(1-0.025, df)
interval <- c(-acc,acc)
names(interval) <- c("Lower limit", "Upper limit")
interval
```

```
## Lower limit Upper limit
##   -1.962691    1.962691
```

Since the calculated t-score (1.136) is within the accept interval, we fail to reject our null hypothesis at a 5% significance level.

(c) Test whether the partial effect of age is different from zero.

We can test whether the partial effect of age is different from zero. We set up our null hypothesis:

$$H_0 : \beta_2 = 0, \beta_3 = 0$$

And the alternative hypothesis:

$$H_1 : \beta_2 \neq 0, \beta_3 \neq 0$$

Since this is a joint hypothesis, an F-test is needed. To do this, we use the Wald-test. The Wald-test is a kind of F-test, that takes heteroskedasticity into account. The F-statistic is calculated using the following formula:

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} * \frac{n - k - 1}{q}$$

```
waldtest(model, vcov=vcovHC(model,type="HC0"), terms=(2:3))
```

```
## Wald test
##
## Model 1: part ~ inc + age + age_sq + educ + you + old + fore
## Model 2: part ~ inc + educ + you + old + fore
##   Res.Df Df      F    Pr(>F)
## 1    864
## 2    866 -2 37.745 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the Wald test we get a p-value < 0.05, and we reject our null hypothesis, and instead accept our alternative hypothesis, H_1, that the partial effect of age is not equal to 0.

2. Set up both a logit and a probit model for participation where you use the described explanatory variables.

The purpose of the probit and logit models are to model the probability of "success" so the models should lie in the unit interval [0,1]. To make this happen the model has a general function G($\cdot$) that takes a value between 0 and 1:

$$P(y = 1|x) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

The function can be written as a logit and probit model.

Logit: $G(z) = \tau(z) = \frac{exp(z)}{[1+exp(z)]}$

Probit: $G(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} exp(-z^2/2)du$

From these models, the partial effect of the independent variables can not be seen directly. The partial effect of a change in a variable can be found by taking the partial derivative of a function g:

$$\frac{\partial P(y = 1|x)}{\partial x_j} = g(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k) \cdot \beta_j$$

Therefore, the estimated partial effects of a roughly continous variable, x_j are:

$$\Delta \hat{P}(y = 1 \mid x) \approx \left[g(\hat{\beta}_0 + x\hat{\beta})\hat{\beta}_j \right] \Delta x_j$$

(a) Estimate the models.

The logit model and the probit model are estimated:

```
model_logit <- glm(part~inc+age+age_sq+educ+you+old+fore, family = binomial(link = "logit"))
model_probit <- glm(part~inc+age+age_sq+educ+you+old+fore, family = binomial(link = "probit"))
screenreg(list( Logit = model_logit, Probit = model_probit), digits = 4)
```

```
##
## ============================================
##                  Logit          Probit
## --------------------------------------------
## (Intercept)       -4.3864 ***    -2.6698 ***
##                   (1.3037)       (0.7756)
## inc               -0.0231 ***    -0.0138 ***
##                   (0.0048)       (0.0028)
## age                0.3295 ***     0.2000 ***
##                   (0.0679)       (0.0402)
## age_sq            -0.0047 ***    -0.0028 ***
##                   (0.0008)       (0.0005)
## educ               0.0386         0.0231
##                   (0.0302)       (0.0181)
## you               -1.1777 ***    -0.7103 ***
##                   (0.1718)       (0.1005)
## old               -0.2354 **     -0.1439 **
##                   (0.0846)       (0.0510)
## fore               1.1908 ***     0.7286 ***
##                   (0.2042)       (0.1214)
## --------------------------------------------
## AIC             1032.1533      1031.6526
## BIC             1070.3196      1069.8189
## Log Likelihood  -508.0766      -507.8263
## Deviance        1016.1533      1015.6526
## Num. obs.        872            872
## ============================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

As explained earlier, the estimates can not be interpreted directly without calculating the partial effect. From this model, it can be seen whether the variables are statisically significant and whether they have a positive or negative effect on the dependent variable.

It can be seen that the income variable has a negative effect in both models while it is also significant at a 1% significance level.

The age variable has a positive effect in both models while the age squared variable has a negative effect in both models. This could indicate, like in 4.1a that the effect becomes negative at a certain age. They are both significant at a 1% significance level.

The education variable is not statistically significant which means that we can not conclude on this variable.

The young kids variable has a negative effect in both models while it is also significant at a 1% significance level.

The old kids variable has a negative effect in both models while it is also significant at a 1% significance level.

The foreign variable has a positive effect in both models while it is also significant at a 1% significance level.

(b) Test whether the partial effect of education is different from zero.

To test whether the partial effect of educ is different from zero in the models, we set up a null hypothesis and an alternative hypothesis:

$$H_0 : \beta_4 = 0$$

$$H_1 : \beta_4 \neq 0$$

These are tested using a t-test, as in 4.1a.

```
screenreg(list(Logit=coeftest(model_logit), Probit=coeftest(model_probit)),digits=4)
```

```
##
## ====================================
##              Logit        Probit
## ------------------------------------
## (Intercept)  -4.3864 ***  -2.6698 ***
##              (1.3037)     (0.7756)
## inc          -0.0231 ***  -0.0138 ***
##              (0.0048)     (0.0028)
## age           0.3295 ***   0.2000 ***
##              (0.0679)     (0.0402)
## age_sq       -0.0047 ***  -0.0028 ***
##              (0.0008)     (0.0005)
## educ          0.0386       0.0231
##              (0.0302)     (0.0181)
## you          -1.1777 ***  -0.7103 ***
##              (0.1718)     (0.1005)
## old          -0.2354 **   -0.1439 **
##              (0.0846)     (0.0510)
## fore          1.1908 ***   0.7286 ***
##              (0.2042)     (0.1214)
## ====================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

Since neither educ is significance for the logit or the probit, we fail to reject H_0. The fact that we fail to reject H_0 does not mean, that we can conclude, that the effect of education is different from zero. Hence, it is not sure, that educ has an effect on participation rate.

(c) Test whether the partial effect of age is different from zero using a likelihood-ratio test.

To perform a hypothesis test on multiple parameters (age and age^2), we can use the likelihood-ratio test.

$$LR = 2(L_{ur} - L_r)$$

The LR-test is similiar to the F-test, but the multiplication of 2 is needed so that the LR approximates a chi-square distribution

We set up our null hypothesis and alternative hypothesis, that is:

$$H_0 : \beta_2 = 0, \beta_3 = 0$$

$$H_1 : \beta_2 \neq 0, \beta_3 \neq 0$$

We already have our logit and probit model, so we only need to setup the logitR and probitR:

```
logitR <- glm(part~inc+educ+you+old+fore, family = binomial(link = "logit"))
probitR <- glm(part~inc+educ+you+old+fore, family = binomial(link = "probit"))
```

First we calculate the LR-score for logit:

```
lrlogit <- 2*(logLik(model_logit)-logLik(logitR))
pvallog <- pchisq(lrlogit,df=2, lower.tail=F)
pvallog
```

```
## 'log Lik.' 2.455753e-14 (df=8)
```

For logit H_0 is rejected since p-value < 0.05, and we instead accept our alternative hypothesis, that the partial effect of age is different from 0.

The same can be done for probit:

```
lrprobit <- 2*(logLik(model_probit)-logLik(probitR))
pvalprob <- pchisq(lrprobit, df=2, lower.tail=F)
pvalprob
```

```
## 'log Lik.' 1.960861e-14 (df=8)
```

The same result can be observed here; p-value < 0.05. Hence, we once again reject H_0 and instead accept our alternative hypothesis: age is different from zero.

3. We want to compare the partial effect of income across the models. Calculate the average partial effect (APE) and comment on the results.

To compare the partial effect of income across the models we use the APE (average partial effect) method on both the probit and logit model and then compare it with the LPM. We do this because it makes little sense to compare the direct coefficient estimates from a logit or probit model (or LPM). APE method is the closest to a comparable estimate between models that we find and it also can be compared with the LPM.

The APE calculates the partial effect for all observations and then takes the average of the effects:

$$\hat{APE} = \hat{\beta}_j \left[n^{-1} \sum_{i=1}^{n} g(x_i \hat{\beta}) \right]$$

This is calculated directly in R:

```
ape_model_logit <- logitmfx(model_logit, data = data, atmean = F)
ape_model_probit <- logitmfx(model_probit, data = data, atmean = F)
screenreg(list(APE_Logit=ape_model_logit, APE_Probit=ape_model_probit, LMP=model),digits=4)
```

```
##
## ============================================================
##                  APE_Logit      APE_Probit     LMP
## ------------------------------------------------------------
## inc                -0.0046 ***    -0.0046 ***   -0.0035 ***
##                    (0.0010)       (0.0010)      (0.0007)
```

```
## age                 0.0657 ***     0.0657 ***    0.0634 ***
##                    (0.0144)       (0.0144)      (0.0129)
## age_sq             -0.0009 ***    -0.0009 ***   -0.0009 ***
##                    (0.0002)       (0.0002)      (0.0002)
## educ                0.0077         0.0077        0.0068
##                    (0.0061)       (0.0061)      (0.0060)
## you                -0.2350 ***    -0.2350 ***   -0.2390 ***
##                    (0.0387)       (0.0387)      (0.0314)
## old                -0.0470 **     -0.0470 **    -0.0475 **
##                    (0.0173)       (0.0173)      (0.0172)
## fore                0.2466 ***     0.2466 ***    0.2572 ***
##                    (0.0403)       (0.0403)      (0.0401)
## (Intercept)                                      -0.3686
##                                                  (0.2530)
## ------------------------------------------------------------
## Num. obs.          872            872           872
## Log Likelihood  -508.0766      -508.0766
## Deviance        1016.1533      1016.1533
## AIC             1032.1533      1032.1533
## BIC             1070.3196      1070.3196
## R^2                                              0.1901
## Adj. R^2                                         0.1836
## ============================================================
## *** p < 0.001; ** p < 0.01; * p < 0.05
```

When looking at the estimate for income in the logit and probit model it can be seen that they are the same and that there is a larger average effect of change in income on the probability of participation than in the LPM. In the logit and probit model the effect of a one unit increase in income is a -0.46% change in the probability of participation, where it in the LPM is -0.35%. The consistency in the results across the models reinforces the robustness of that an increase in income is associated with a lower probability of participation.

4. We want to compare the partial effect of the foreign variable across the models. Calculate the Average Partial Effect (APE) and comment on the results.

APE is used to to examine a dummy variable, foreign. We calculate APE using the method for discrete variables.

$$\hat{APE} = n^{-1} \sum_{i=1}^{n} [G[\beta_0 + \beta_1 income + \beta_2 age + \beta_3 agesq + \beta_4 educ + \beta_5 youngkids + \beta_6 oldkids + \beta_7] - G[\beta_0 + \beta_1 income + \beta_2 age + \beta_3 agesq$$

So the APE when foreign (β_7) is included minus APE when foreign is excluded. For each observation (i), the estimated effect of changing foreign from 0 to 1 is calculated.

We start by setting up a matrix for all variables and define if a person is foreign(1) or not(0):

```
cdata <- cbind(1, as.matrix(data[,c("income","age","agesq","educ","youngkids","oldkids","foreign")]))
cdata1 <- cdata
cdata1[,8] <- 1
cdata2<-cdata
cdata2[,8] <- 0
```

Afterwards, the coefficients for both models is found

```
pcoef <- model_probit$coefficients
lcoef <- model_logit$coefficients
```

Now, APE can be calculated for probit and logit. Starting with probit:

```
probitAPE <- mean(pnorm(cdata1%*%pcoef)-pnorm(cdata2%*%pcoef))
probitAPE
```

```
## [1] 0.2493686
```

And for logit:

```
logitAPE <- mean(pnorm(cdata1%*%lcoef)-pnorm(cdata2%*%lcoef))
logitAPE
```

```
## [1] 0.3362055
```

It can be seen, that the probit APE is approximately 0.25, which is almost the same as for the LPM (0.26). However, for logit, it can be seen, that there is a quite big difference, as logit gives a result of 0.336, which indicates, that the partial effects for probit is almost the same as for LPM, where as for logit there is a difference. Which one is more correct, is hard to interpret.

5. Why is the Average Partial Effect (APE) preferred over the Partial Effect at the Average (PEA)?

The partial effect at the average (PEA) can be seen as the partial effect at the "average" of a variable. It is written like this:

$$P\hat{E}A_j = g(\bar{x}\hat{\beta})\hat{\beta}_j$$

A challenge can arise when using the PEA for a variable that can only take a value of 0 and 1. An example could be the foreign variable used earlier in this assignment. When a variable has a value of either 0 or 1 it would not make sense to calculate the partial effect of the average person as the PEA does.

On the other hand, the average partial effect (APE) calculates the partial effect for all observations and then takes the average of the effect. As shown earlier in this assignment, it is written like this:

$$A\hat{P}E = \hat{\beta}_j[n^{-1}\sum_{i=0}^{n} g(x_i\hat{\beta})]$$

Therefore, the reason that APE is preferred over PEA is that the APE shows the partial effect where as the PEA shows the partial effect of the average person.

6. Compare the models' predictive abilities by calculating the percent correctly predicted for each model.

Using a goodness of fit estimate called percently correctly predicted (PCP), which shows how big a share of the data that the models predicts correctly. However, PCP is not necessarily always a good indicator of the model. E.g. if 300 answers yes and 60 answers no, but the model predicts that everyone answers yes, it will have a high PCP, but is not necessarily a good predictor. To examine when the model predicts correctly, we set up a probability limit for fitted values of the models:

$$\tilde{y} = 1, if\ \tilde{y}\ \geq\ c$$

$$\tilde{y} = 0, if\ \tilde{y}\ \leq\ c$$

The probability limit is set to 0.5. Hence, fitted values of the three models below 0.5 will predict that the person not participating in the labor force (y=0), while fitted values over 0.5 will predict, that the person participates in the labor force (y=1).

PCP is calculated in R in the following way:

```
y <- data$participation
lmp_pcp <- 100 * mean((model$fitted.values > 0.5) == y)
logit_pcp <- 100 * mean((model_logit$fitted.values > 0.5) == y)
probit_pcp <- 100 * mean((model_probit$fitted.values > 0.5) == y)
PCP <- c(lmp_pcp, logit_pcp, probit_pcp)
names(PCP) <- c("LMP PCP", "Logit PCP", "Probit PCP")
PCP
```

```
##    LMP PCP  Logit PCP Probit PCP
##   67.43119   68.11927   68.11927
```

It can be seen that the three models are almost equally good at predicting. The logit and probit models only predict marginally better at 68.12%, while the LPM correctly predicts 67.43%.