**Tutorial Nr. 8 – Linear Probability Model**

**Timing**

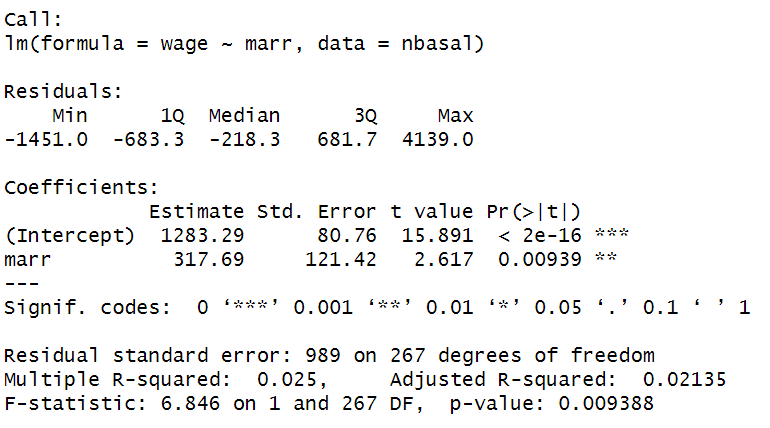
* particify (5min)
* questions & revision (15min)
* exercise (40min)
* discussion (30min)

**Questions?**

**Revision**

1. **Binary variables**

* What is a binary variable, also called dummy variable?
  + Categorical variable w/ two categories, 2 possible values: either 0 or 1
* If we want to measure the effect of e.g. being married on wage 🡪 we could run a regression w/a variable “married” & estimate its marginal effect on “wage”



* + What is a ref cat? And what would be the reference category here?
    - The cat you compare to; var=0
  + How to interpret the coefficient of “married”?
* What would be another alternative to get the difference between wages of those two groups?
  + Divide sample into married & non-married, compare their avg wages
* Lecture graph:
  + Forget about b1xi:
    - what is the average balance of a homeowner? (b0+d)
    - what is the average balance of someone not owning a home? (b0)
    - What is the difference? (d)

1. **Binary variables as endogenous variable: LPM**

* Between which values do the observations of our dependent variable vary?
  + 0 or1 🡪 observe either a 0 or a 1; e.g. decision to go to university
* How does the left side of our regression function change?
  + before measure effect on y, now on probability of y=1
* How does the interpretation change?
  + Before: change in exogenous var by 1 unit associated w/ change in y by 1 unit
  + Now: change in exogenous var by 1 unit associated w/ change in probability of y = 1
* Which method do we apply for the estimation?
  + Still ordinary least squares
* What are the problems of the LPM?
  + Probabilities >1 and <0 🡪 doesn’t make sense
  + Bi-modal distribution of residuals 🡪 residuals do not have a constant variance 🡪 heteroskedasticity

**Questions?**

**Solution exercise**

1. **Linear Probability model**

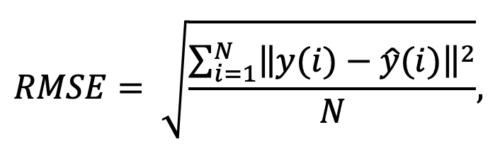
* Create dummy variable
* Model interpretation
  + Who wants to present? How did you choose the variables in your model?
  + Interpretation:
    - Intercept: probability Airbnb has highrating if all other var = 0
      * Coeff\*100% 🡪 4.958e-01 = 4.958\*(10^-1) = 0.4958 🡪 49.58%
    - nrofreviews:
      * -2.158e-03 = -2.158\*10^-3 = -0.002158 🡪 0.2 pp
      * one more review is associated with an on average decrease in prob of high rating of 0.2 pp
    - Price:
      * 2.372e-04 = 2.372\*10^-4 = 0.0002372 🡪 0.02 pp
      * increasing price by 1$, increases prob of high rating by 0.02 pp
    - d\_gym:
      * 5.809e-02 = 5.809 \* 10^-2 = 0.05809🡪 5.8 pp
      * having a gym in the Airbnb is associated w/ an avg increase in prob of being high rated by 5.8 pp

🡪 all significant at the 0.001 level

* + alternative; just to see “.” 🡪 includes all exogenous variables
    - why problematic? On model selection:
      * Coefficients can be significant, but meaningless; especially in large data sets more likely to get significant results
      * Should start w/ theoretical considerations & RW 🡪 Which variables we include in our model will always depend on the RQ
      * Increasing nr of EV always increases R2, but higher risk of multicollinearity (strong correlation between var 🡪 makes estimation more inprecise, bc if both equally explain variation in y, cannot detect where variation comes from
    - R2 not used for binary models 🡪 y only varies between 0 and 1
  + Why LPM problematic?
    - Can predict probabilities <0; >1
    - Assumptions required for OLS might not be met (e.g. homoskedasticity)

1. **RMSE**

* Also: standard deviation of residuals
* measures how well the model predicts our target value 🡪 accuracy measure
* average distance between actual & predicted values

  
🡪 measures how dispersed our residuals are around the fitted regression line

🡪 magnitude of unexplained variation

* always interpreted based on the scale of our dependent variable
  + 0 🡪 model fits data perfectly
  + The lower the better
  + RSME = 4; average difference between actual & predicted value = 4

**LPM RMSE:**

* + 0.49 🡪quite high again bc ranging from 0 to 1
  + Can also be found in output: *residual standard error*

**Lin Reg RMSE & R2**

* What does R2 tell?
* Compare R2 of mod1 and mod2
  + Mod1: 0.01041 🡪 only 1% explained
  + Mod2: 0.43 🡪 model explains 43% of the variation in y; quite good; better than mod1
* RMSE
  + RMSE=59, which is moderately high given that the price ranges from 10-986; it means that, on average, the model's predictions are off the actual values by about 59 units