**Tutorial Nr. 7 – Linear Regression**

**Timing**

* announcements + particify (5min)
* questions & revision (15min)
* time for exercise (45min)
* discussion (20min)

**Prepare**:

* particity link + jupyterhub link
* open lecture slides
* one tutor as co-host
* breakout rooms à 40

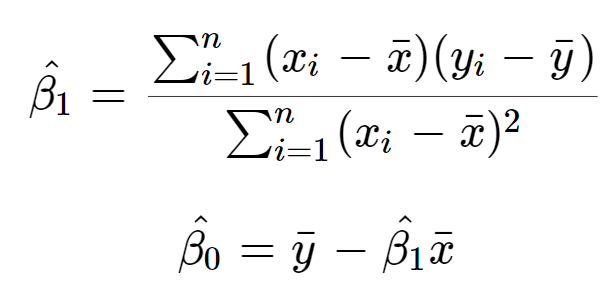
**Questions?**

**Revision**

*Bis wohin in VL gekommen?*

1. **Basic idea of linear regression**

* What is the difference between beta1 and beta1 hat?
  + True value/relationship vs estimated 🡪 meaning population ground truth vs relationship we estimate by taking a sample of the population 🡪 will never be able to model the truth
* What are residuals?
  + Difference between observation of y & estimated value of y
* What is the basic idea of OLS?
  + Min RSS 🡪 the difference between obs & est 🡪 getting as close as possible to true relationship
* If we derive RSS to beta 1. What does the formula of beta 1 estimator consists of?
  + cov/var



1. **Unbiasedness & Efficiency**
   * What are the 2 criteria by which we choose our estimator?
   * What do they mean?
     + Unbiasedness: expected value in population = estimator; estimator should not vary systematically from the true value
     + Efficiency: get most precise estimator, i.e. lowest variance
2. **Goodness of fit & inference**

* How can we evaluate our model? What does it mean?
  + R2: share of variation in y that is explained by the model/variation in x
* How can we assess whether there is a systematic significant relationship between x & y?
  + T-test/p-value 🡪 reject the H0 at certain significance level; mostly 5%
* What is our H0?
  + No system relationship, i.e. beta1=0

1. **Interpretation**
   * What is beta0, what is beta1?
     + Intercept + slope
   * How to interpret beta1, beta0
     + Beta0 = mean if all other var=0
     + Beta1: changing x by 1 unit increases y by 1 unit 🡪 marginal effect

**Questions?**

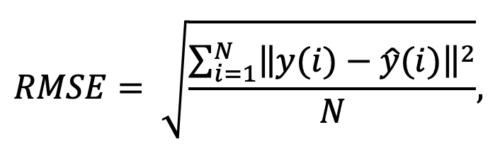
**Solution exercise**

1. descriptives
   * nr of obs = nr of rows or describe command
   * nr of variables = nr of columns or str
   * variable types (needed for interpreting it + values of variables)
   * with the “$” you only select one variable
   * describe also useful for missing values
   * help(airbnbsmall) 🡪 provides information about the data set & variable description
2. Linear Regression interpretation

* Intercept: mean price if all other var are zero: 18$ probably per night
* d\_breakfast: if Airbnb has breakfast expected price decreases by 11$ c.p.
* n\_accommodates: for every further accommodate, expected price increases by 25$ c.p.
* n\_nrofreviews: for every further review, expected price decreases by 0.11$ c.p.

🡪 all significant at the 0.001 level

* R2: model explains 43% of the variation in y; quite good
* **RMSE (root mean squared error)**: measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value (accuracy).



* + Predict the values of model; deduct them from observed y values; take root mean square
  + RMSE=59, which is quite high given that the price ranges from 10-986; it means that, on average, the model's predictions are off by about 59 units
* Will stay in the room 5min longer for questions
* Save folder locally

---

1. LPM interpretation
   * Probability that high rating if all other var=0: 0.05%
   * N\_nrofreviews: one more review is associated with an on average decrease in prob of high rating of 0.022 pp
   * Price: increasing price by 1$, increases prob of high rating by 0.024 percentage points
   * d\_gym: having a gym in the Airbnb is associated w/ an avg increase in prob of being high rated by 0.058 pp

🡪 all significant at the 0.001 level

* + RMSE of 0.49 quite high again bc ranging from 0 to 1
  + R2 cannot be used as explaining variation in binary outcome (either 0 or 1) questionable
  + If we look at sum(in\_predictions) we see that there are negative probabilities