Simple Linear Regression and OLS

Contents and Goals

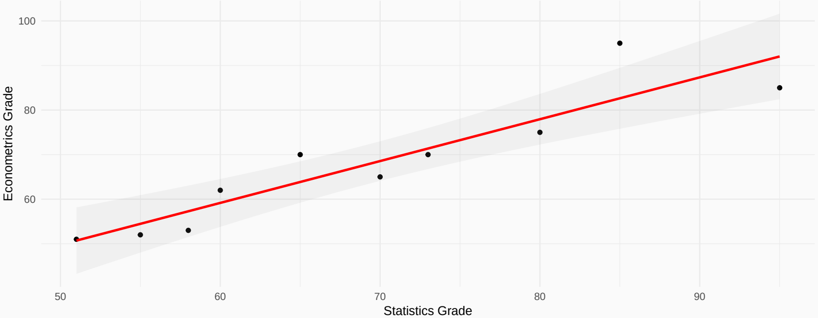
- Get to know the basic features of probability distributions
    - Mean / Expected Value
    - Variance and Standard Deviation
    - Covariance and Correlation
    - Standardisation
- Learn how they work
- Learn how we can manipulate them

Linear Regression

- Predict one variable from another: One simple way to do so is using linear regression (SLR)
- SLR is about predicting a dependent variable (*regressand*) from one independent variable (*regressor*).
- A SLR model is given by:

  $$y = \beta_0 + \beta_1 x + \epsilon$$

  where y is the regressand and x the regressor. ϵ denotes the error of the model, i.e. the residuals. Regression is often used to model real-world relationships.
- Note: Nothing else than any linear function $y = mx + b$.
  - In regression, β_0 is b, and β_1 is m, the slope.
- The regression line is the one line that crosses or passes the observation with the smallest amount of squared error.



- Line does not cross a single data point, but it minimises the squared deviations of all the single points from the line. - Deviations are called "residuals" and in a linear regression model, they sum up to zero. - Looking at the residuals:

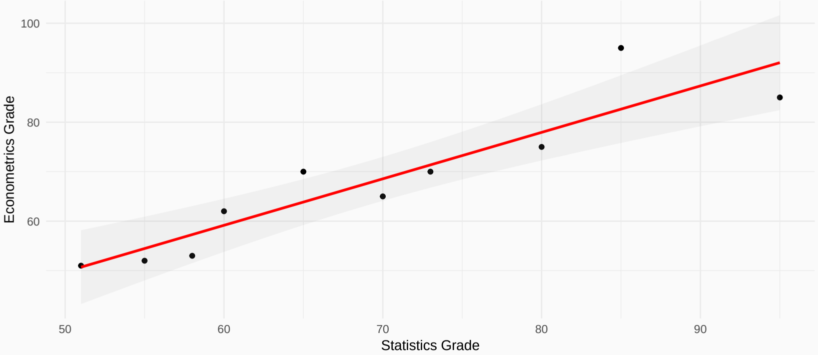
```
##          1          2          3          4          5
## -7.0271078 12.3632441 -2.9415800 -1.3683337 -3.5512281
##          7          8          9         10
##  2.8391237 -4.2828059 -2.4657004  0.2904404
```

Interpretation by example

- In the example, we have ten students who took exams in Statistics and exams in Econometrics. We know their grades in Statistics. Now, we would like to predict the Econometrics grade of any other student who took the Statistics exam.
- Model:

$$\widehat{econgrade} = \hat{\beta}_0 + \hat{\beta}_1 statsgrade + \hat{\epsilon}$$

- The hats ˆ are used because we look at empirical data, and we can only *estimate* a model, but not claim that we have an image of the theoretical model here



```
summary(lm(y~x, data=grades_df))$coefficients
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 2.8187652 10.0801739 0.2796346 0.7868516450
## x           0.9390352  0.1429934 6.5669819 0.0001753288
```

- Clear linear relationship between the students' statistics grades and the econometrics grades.

- Essential problem with this interpretation: Do we have a causal relationship or just a correlation? Do we know that the line we drew is the best guess for a prediction?

Ordinary least squares

- Regression line = line that fits our points so that it minimises the squared deviations (residuals) of the points from the line.
- To construct such a line, we need to know the intercept and the slope coefficient(s) for x. The formulas are quite simple:

$$\hat{\beta}_1 = \frac{Cov[x, y]}{Var[x]}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

- Proof by a formal minimization problem, we want to minimize $\hat{\epsilon}^2$.

—

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2$$

- Minimise by taking the first order conditions (FOCs) by $\hat{\beta}_0$ and

- Now derive by $\hat{\beta}_1$:

$$
\begin{aligned}
& 2 \times \sum_{i=1}^{n} -x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \overset{!}{=} 0 \mid \text{:(-2)} \\
\iff & \sum_{i=1}^{n} x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \overset{!}{=} 0 \\
\iff & \sum_{i=1}^{n} x_i(y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i) \overset{!}{=} 0 \\
\iff & \sum_{i=1}^{n} x_i(y_i - \bar{y}) + \sum_{i=1}^{n} x_i(-\hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) \overset{!}{=} 0 \\
\iff & \sum_{i=1}^{n} x_i(y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) \overset{!}{=} 0 \\
\iff & \sum_{i=1}^{n} x_i(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) \mid \text{counterintuitive but possible:} \\
\iff & \sum^{n}(x_i - \bar{x})(y_i - \bar{y}) = \hat{\beta}_1 \sum^{n}(x_i - \bar{x})^2
\end{aligned}
$$

Assumptions

Before interpreting SLR models, one should be aware of its assumptions, namely: 1. The population model is linear in its parameters: $y = \beta_0 + \beta_1 x + \epsilon$ 2. The sample at hand is a random sample from the population model.
3. There is variation in the x_i: $\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$ 4. $\mathbb{E}\left[\epsilon | x\right] = 0$ and therefore $\mathbb{E}\left[\epsilon_i | x_i\right] = 0$ 5. Homoskedasticity (Equality of variances) is given: $Var\left[\epsilon | x\right] = Var\left[\epsilon_i | x_i\right] = \sigma^2$

Note that this only holds as long as the the assumption of independence is met, i.e. $\frac{\delta \epsilon}{\delta x} = 0$