

## 1. Problem

**Theory:** Consider a linear regression of  $y$  on  $x$ . It is usually estimated with which estimation technique (three-letter abbreviation)?

This estimator yields the best linear unbiased estimator (BLUE) under the assumptions of the Gauss-Markov theorem. Which of the following properties are required for the errors of the linear regression model under these assumptions?

independent / zero expectation / normally distributed / identically distributed / homoscedastic

**Application:** Using the data provided in `linreg.csv` estimate a linear regression of  $y$  on  $x$ . What are the estimated parameters?

Intercept:

Slope:

In terms of significance at 5% level:

$x$  and  $y$  are not significantly correlated /  $y$  increases significantly with  $x$  /  $y$  decreases significantly with  $x$

**Interpretation:** Consider various diagnostic plots for the fitted linear regression model. Do you think the assumptions of the Gauss-Markov theorem are fulfilled? What are the consequences?

**Code:** Please upload your code script that reads the data, fits the regression model, extracts the quantities of interest, and generates the diagnostic plots.

## Solution

**Theory:** Linear regression models are typically estimated by ordinary least squares (OLS). The Gauss-Markov theorem establishes certain optimality properties: Namely, if the errors have expectation zero, constant variance (homoscedastic), no autocorrelation and the regressors are exogenous and not linearly dependent, the OLS estimator is the best linear unbiased estimator (BLUE).

**Application:** The estimated coefficients along with their significances are reported in the summary of the fitted regression model, showing that  $x$  and  $y$  are not significantly correlated (at 5% level).

Call:

```
lm(formula = y ~ x, data = d)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.55258	-0.15907	-0.02757	0.15782	0.74504

Coefficients:

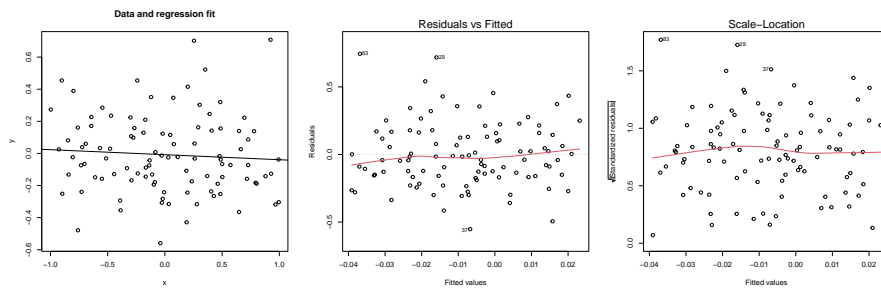
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.007988	0.024256	-0.329	0.743
x	-0.031263	0.045420	-0.688	0.493

Residual standard error: 0.2425 on 98 degrees of freedom

Multiple R-squared: 0.004811, Adjusted R-squared: -0.005344

F-statistic: 0.4738 on 1 and 98 DF, p-value: 0.4929

**Interpretation:** Considering the visualization of the data along with the diagnostic plots suggests that the assumptions of the Gauss-Markov theorem are reasonably well fulfilled.



**Code:** The analysis can be replicated in R using the following code.

```
## data
d <- read.csv("linreg.csv")
## regression
m <- lm(y ~ x, data = d)
summary(m)
## visualization
plot(y ~ x, data = d)
abline(m)
## diagnostic plots
plot(m)
```