

Econometrics , Mock Exam

Time: 75 min

1. Consider a dataset which includes information about houses sold in a particular city. The dataset contains the following variables:

- Y : Sale price in US dollars for which the house was sold.
- X_1 : Size of the house in square feet.
- D_2, D_3, D_4, D_5 : Dummy variables representing the location of the house, with D_1 as the baseline category. Specifically, $D_2 = 1$ if the house is in location 2 and 0 otherwise, $D_3 = 1$ if in location 3 and 0 otherwise, $D_4 = 1$ if in location 4 and 0 otherwise, and $D_5 = 1$ if in location 5 and 0 otherwise. D_1 (the most desirable location) is excluded from the model.

The following regression model is considered to analyze the relationship between the sale price of the houses and the explanatory variables:

$$\log(Y) = \beta_0 + \beta_1 X_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 D_4 + \beta_5 D_5 + \epsilon \quad (1)$$

where ϵ is the error term with $\sim N(0, \sigma^2)$. The results (and some accompanying plots and tests) can be found below.

Call:

```
lm(formula = log(price) ~ size + D_2 + D_3 + D_4 + D_5, data = houses)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.45805	-0.13208	-0.01486	0.15261	0.58089

Coefficients:

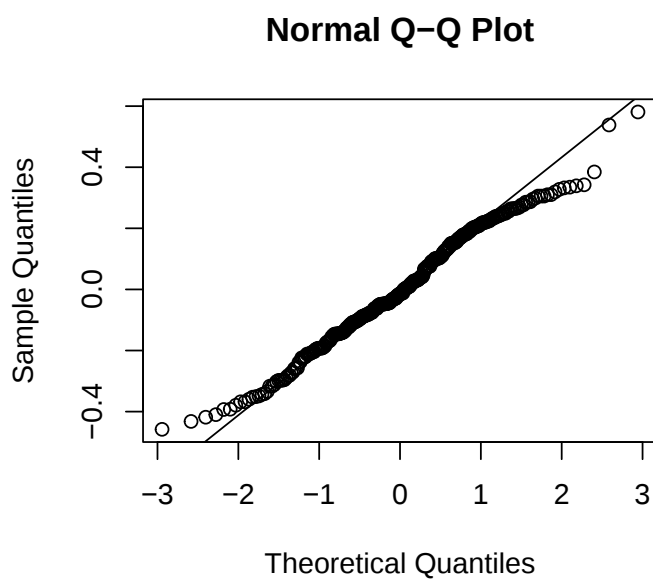
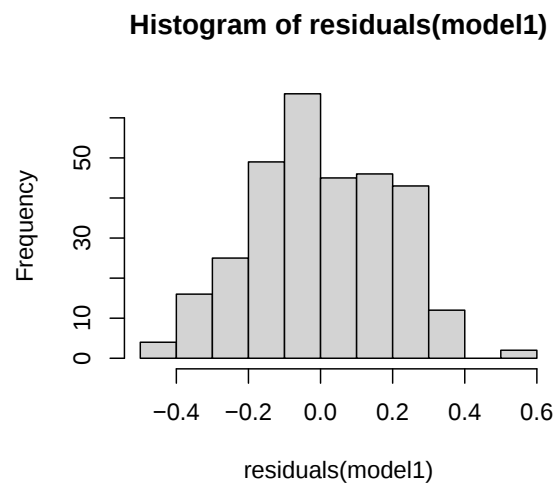
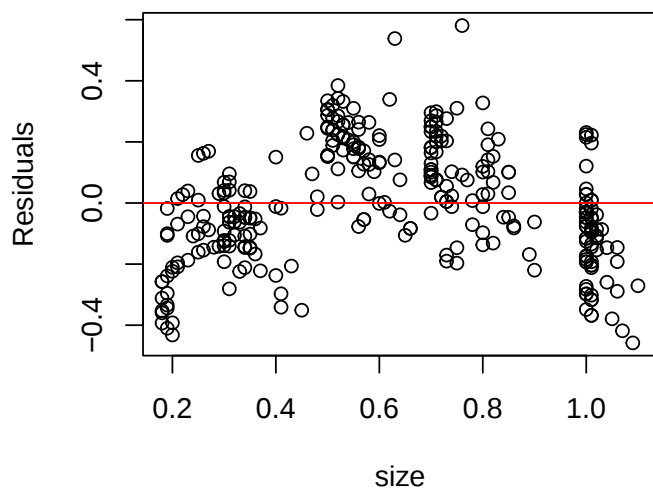
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.7293962	0.0331951	262.973	< 2e-16 ***
size	0.0288300	0.0004341	66.407	< 2e-16 ***
D_2	-0.0130551	0.0379606	-0.344	0.731151
D_3	-0.1528734	0.0426772	-3.582	0.000397 ***
D_4	0.1199752	0.0410927	2.920	0.003768 **
D_5	0.0055677	0.0384868	0.145	0.885072

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.1924 on 302 degrees of freedom

Multiple R-squared: 0.9451, Adjusted R-squared: 0.9442

F-statistic: 1039 on 5 and 302 DF, p-value: < 2.2e-16



Jarque Bera Test
data: residuals(model_diamonds)
X-squared = 2.4242, df = 2, p-value = 0.2976

studentized Breusch-Pagan test
data: model1
BP = 4.0731, df = 5, p-value = 0.5389

- a) What are the estimates for the coefficients β_1 and β_3 ? Report the values and provide interpretations.
- b) Does the assumption of exogeneity seem to be fulfilled in this example? Justify your answer!
- c) Does the assumption of a homoscedastic error term ϵ seem to be fulfilled? Justify your answer!
- d) Does the assumption of normality of the error term ϵ seem to be fulfilled? Justify your answer!

2. We now estimate a new model, where we add size squared (X_1^2) as a new predictor, i.e.,

$$\log(Y) = \gamma_0 + \gamma_1 X_1 + \gamma_2 X_1^2 + \gamma_3 D_2 + \gamma_4 D_3 + \gamma_5 D_4 + \gamma_6 D_5 + \varepsilon \quad (2)$$

Regression results are given below:

```
Call:
lm(formula = log(price) ~ size + I(size^2) + D_2 + D_3 + D_4 +
    D_5, data = houses)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.41667	-0.09137	0.00812	0.08983	0.38151

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	8.113e+00	4.233e-02	191.658	< 2e-16 ***
size	5.634e-02	1.602e-03	35.160	< 2e-16 ***
I(size^2)	-2.130e-04	1.218e-05	-17.492	< 2e-16 ***
D_2	-1.373e-01	2.770e-02	-4.955	1.21e-06 ***
D_3	-2.537e-01	3.065e-02	-8.277	4.15e-15 ***
D_4	-2.814e-02	3.020e-02	-0.932	0.352
D_5	-1.242e-01	2.814e-02	-4.414	1.42e-05 ***

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.1357 on 301 degrees of freedom

Multiple R-squared: 0.9728, Adjusted R-squared: 0.9722

F-statistic: 1792 on 6 and 301 DF, p-value: < 2.2e-16

- In model (2) (i.e. the model with the additional regressor X_1^2), what is the expected $\log(\text{price})$ ($\log(Y)$) of a house that is $120m^2$ and a location of 3?
- In model (2), calculate the vertex point of the estimated regression parabola and give an interpretation of what this means in our problem.
- Should the quadratic term of `size` be included? Justify your answer.
- Should variable `size` be included to the model at all? Justify your answer.

3. True or false? Where is the error in a false statement? The question is considered to be answered correctly if you classify a true statement as true or if you correct a false statement such that it becomes true.
- a) The coefficient β_2 in the regression $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \dots$ is the slope of Y with respect to X_2 .
 - b) Consider the case where we use a dummy variable D in the regression: $Y = \beta_0 + \beta_1 X_1 + \beta_2 D$. In this case, we have two different regression lines with slopes differing by β_2 .
 - c) Non-normality of errors can be sometimes overcome by transformations.
 - d) If the p -value from an Augmented Dickey-Fuller (ADF) test is less than the chosen significance level, we reject the null hypothesis and conclude that the time series is non-stationary.
 - e) The Ljung-Box test checks if the residuals are white noise, indicating no autocorrelation.
 - f) The omitted variable bias depends on the correlation between included and omitted regressors.
 - e) For an AR(2) model, the empirical autocorrelation function (ACF) is expected to cut off at the lag 2.
 - f) If a regression yields residuals which follows AR(1) model, then the standard errors of coefficients in the estimated model are still unbiased.
 - g) Invertibility of a MA process requires the roots of characteristic polynomial to lie in the unit circle.

4. Consider an autoregressive model of order 1, AR(1), defined by the equation:

$$Y_t = \phi Y_{t-1} + \epsilon_t$$

where Y_t is the time series value at time t , ϕ is the autoregressive parameter, and ϵ_t is a white noise error term with variance σ^2 .

- a) Provide the condition for the stationarity of the process Y

- b) For a stationary process Y , compute the mean $E[Y_t]$ and the variance $Var(Y_t)$.