FORK 1002 Statistics
Exercise Set 2: Simple Regression

1. The dataset *houses_sample.xls* contains information about the sales of 20 houses and apartments in 2010 in an area of Oslo:

$m2$:	Size measured in square meters
$Rooms$:	Number of rooms
$Priceindication$:	Price indication (in thousands of NOK) before sale
$Salesprice$:	Sales price (in thousands of NOK)
$Salespriceincldebt$:	Sales price + debt of house/apartment
$Debt$:	Salespriceincldebt − Salesprice

   (a) What kind of dataset is this?

   (b) On which level of measurement is each variable?

2. Let Y_i denote the sales price and let X_i denote the indicated price before sales. In a preliminary set of computations, the following is obtained:

$$\sum_{i=1}^{n} Y_i = 34\ 915, \quad \sum_{i=1}^{n} X_i = 33\ 390, \quad \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = 5\ 905\ 314,$$

$$\sum_{i=1}^{n} (X_i - \overline{X})^2 = 5\ 269\ 895, \quad \sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X}) = 5\ 467\ 158,$$

   (a) Compute the sample means of sales price and price indication

   (b) Compute their sample variances

   (c) Compute their sample covariance

   (d) Compute their sample correlation

   (e) In order to study the relationship between sales and indicated prices further, the following model is estimated:

$$Y_i = B_0 + B_1 X_i + u_i. \tag{1}$$

   The estimated model is denoted

$$\widehat{Y}_i = \widehat{B}_0 + \widehat{B}_1 X_i, \tag{2}$$

   where $\widehat{B}_0$ and $\widehat{B}_1$ are estimates of B_0 and B_1. Show that the estimates are equal to 13.8107 and 1.0374, respectively.

(f) Interpret the estimates

(g) What are the predicted sales prices according to the estimated model for the following price indications?: 500 thousand Norwegian kroner, 1 million kroner, 2 million kroner and 5 million kroner

(h) For observation number 11 the indicated sales price is 2 690 and the sales price is 3 010. What is the prediction error or residual value of the estimated model for observation number 11?

(i) It turns out that the explained variation (Explained Sum of Squares; ESS) of the estimated model (2) is equal to 5 671 804. What is the unexplained variation (Residual Sum of Squares; RSS)?

(j) Estimate and interpret the R-squared of (2)

(k) Estimate the adjusted R-squared of (2)

(l) Estimate the standard error of regression of (2)

3. In this exercise the data from exercise 1 and the results from exercise 2 above will be used:

   (a) The standard error of $\widehat{B}_0$, that is, $se(\widehat{B}_0)$, is equal to 86.6596. Test if the intercept or constant B_0 is significantly different from 0 at the 5% significance level

   (b) Estimate a 95% confidence interval for B_0, and provide an interpretation

   (c) The standard error of $\widehat{B}_1$, that is, $se(\widehat{B}_1)$, is equal to 0.0496. Test if B_1 is significantly *bigger* than 1 at a 1% significance level

   (d) Estimate a 99% confidence interval of B_1

4. In this exercise the data from exercise 1 above will be used:

   (a) The estimate $\widehat{B}_0$ of B_0 in the model $Y_i - X_i = B_0 + u_i$ is equal to 76.25. Give an interpretation of the estimate

   (b) The standard error of $\widehat{B}_0$, that is, $se(\widehat{B}_0)$, is equal to 25.1780. Test if the real estate agents on average have a tendency to provide an indicative price that is lower than the sales price (use a 1% significance level)

5. Computer exercises:

   (a) Load *houses_sample.xls*. [Hint for SPSS: File → Open → Data...]

   (b) Make a scatterplot of the variables Y_i and X_i [Hint for SPSS: Graphs → Chart Builder..., choose "Scatter/Dot" in the "Gallery" tab, drag the chosen scatter type (e.g. *Simple Scatter*) into the "Chart preview..." area, choose the variables, press "OK"]

   (c) Compute the correlation between X_i and Y_i [Hint for SPSS: Analyze → Correlate → Bivariate..., put "*Salesprice*" and "*Priceindication*" into the "Variables" box, tick "Pearson's", press "OK"]

(d) Estimate the model $Y_i = B_0 + B_1 X_i + u_i$. [Hint for SPSS: Analyze → Regression → Linear, choose "Salesprice" as dependent variable, choose "Priceindication" as independent variable, press "OK"]

(e) Compute a 95% confidence interval for B_0 [Hint for SPSS: Analyze → Regression → Linear, press "Statistics", tick "Confidence intervals" and choose the confidence level, press "Continue"]

(f) Compute a 99% confidence interval for B_1 [Hint for SPSS: Similar to previous]

(g) Make a new variable that is equal to the predicted salesprice of the estimated model [Hint for SPSS: Analyze → Regression → Linear, press "Save", tick "Unstandardized" in the "Predicted Values" area, press "Continue", press "OK"]

(h) Make a new variable that is equal to the residuals of the estimated model [Hint for SPSS: Similar to previous, but tick "Unstandardized" in the "Residuals" area instead]

(i) Make a new variable called "$pricediff$" equal to $Y_i - X_i$ [Hint for SPSS: Transform → Compute Variable..., write "$pricediff$" in the "Targe Variable" box, write "Salesprice - Priceindication" in the "Numeric Expression" box, press "OK"]

(j) Estimate the model $Y_i = B_1 X_i + u_i$ (that is, no constant) [Hint for SPSS: Analyze → Regression → Linear, press "Options", untick "Include constant in equation", press "OK", choose "Salesprice" as dependent variable, choose "Priceindication" as independent variable, press "OK"]

(k) Estimate the model in 4(b), that is, $Y_i - X_i = B_0 + u_i$ [Hint for SPSS: Make a constant variable equal to 1 and call it, say, `const`. Then estimate the model with `const` as the only independent variable while unticking "Include constant in equation" via "Options"]