Tutorial: Simple Linear Regression

Open the dataset `hospitaldata.dta`.

Exercises:

1. Calculate the Pearson correlation for the percent of patients who say their nurse always communicated well (nursealways) and the percent of patients who would always recommend the hospital (recommendyes).

   ```
   pwcorr recommendyes nursealways, sig
   ```

   These two variables are correlated. However, simple linear regression gives us a more intuitive measure of the relationship between the two variables. Specifically, we can state: ”For a one percent increase in the percent of patients who say their nurse always communicated well, we would, on average, expect to see a corresponding increase of B% of patients who would always recommend the hospital.” Here B is determined by fitting an appropriate linear regression model.

2. Now that you have established that these variables are correlated, you decide to fit a linear regression model to assess the relationship between `recommendyes` and `nursealways`. State your model.

   $Yi =$ percent of patients who always recommend the hospital
   $Xi =$ perecnt of patients who say that the nurse always communicates well

   $$Y_i = \alpha + \beta X_i + \epsilon_i$$

   where $\epsilon_i \sim N(0, \sigma^2)$. Equivalently, we could write:

   $$\mu_{yi} = E(Y_i|X_i) = \alpha + \beta X_i$$

   where $Y_i \sim N(\mu_{yi}, \sigma^2)$.

   Goal is to estimate and obtain measures of uncertainty for α and β. We use the method of least squares for estimation.

3. Construct a scatter plot with `nursealways` on the x-axis and `recommendyes` on the y axis. Use the scatterplot to evaluate the assumptions of simple linear regression.

   ```
   twoway (scatter recommendyes nursealways)
   ```

   Assumptions:

   - Independent observations

- $Y|X$ is normally distributed
- Homoscedasticity (constant variance)
- Linearity

4. Fit the linear regression model. Provide estimates, confidence intervals, and interpretations of the regression coefficients α and β.

```
. regress recommendyes nursealways

      Source |       SS       df       MS              Number of obs =    3570
-------------+------------------------------           F(  1,  3568) = 2723.72
       Model |  144368.851     1  144368.851           Prob > F      =  0.0000
    Residual |  189118.972  3568  53.0041962           R-squared     =  0.4329
-------------+------------------------------           Adj R-squared =  0.4327
       Total |  333487.823  3569  93.4401297           Root MSE      =  7.2804

------------------------------------------------------------------------------
recommendyes |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 nursealways |   1.159487   .0222169    52.19   0.000     1.115928    1.203046
       _cons |  -19.21559   1.712829   -11.22   0.000    -22.57381   -15.85737
------------------------------------------------------------------------------
```

Our fitted regression line is:

$$Y_i = -19.2 + 1.16X_i + \epsilon_i$$

where $\epsilon \sim N(0, 7.3^2)$.

Confidence intervals for α and β, respectively, are (-22.57, -15.86) and (1.12, 1.2).

For a 1% increase in patients reporting their nurse communicated well, β is corresponding average increase in the percent of patients who would always recommend the hospital 1.16%.

α is the mean value of the response Y_i when $X_i = 0$ and for this example has no meaningful interpretation. (However, it is necessary for constructing the regression line and making subsequent predictions).

5. Test the hypothesis that $H_0 : \beta = 0$ versus the alternative that $H_A : \beta \neq 0$.

We find that $\hat{\beta} = 1.16$, $\hat{se}(\hat{\beta}) = 0.02$, and $t = 52.2$. Under H_0, $t = \hat{\beta}/\hat{se}(\hat{\beta}) \sim t_{3570-2}$, and our p-value < 0.0001. Therefore, we reject the null hypothesis and conclude that the percent of patients who say a nurse always communicates well is positively correlated with the percent of patients who would always recommend a hospital.

6. What is the value of R^2. Interpret this quantity.

0.433

43% of the variability among the observed values of `recommendyes` is explained by the linear relationship with `nursealways`.

7. Examine a residual plot. Using R^2 and the plot, does the model appear to fit well? (Are there any outliers?)

```
rvfplot
rvpplot nursealways
```

   We don't see any strong trends or outliers in the residual plots.

8. Using the regression line, predict the expected percent of patients who always recommend the hospital when the reported percent of nurses who always communicate well is 80%? Construct corresponding 95% confidence interval.

   Denote $\bar{Y}^{80}$ as the predicted average percent of patients who always recommend a hospital among hospitals with patients reporting that nurses always communicate well 80% of the time.

   $\bar{Y}^{80} = -19.2 + 1.16 * 80 = 73.6$

```
. lincom _cons + 80*nursealways

 ( 1)  80*nursealways + _cons = 0

------------------------------------------------------------------------------
recommendyes |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   73.54339   .1399631   525.45   0.000     73.26897     73.8178
------------------------------------------------------------------------------
```

   A 95% confidence interval for $\bar{Y}^{80}$ is (73.2690, 73.8178).

9. For a new hospital with 80% of patients reporting that nurses always communicate well, predict the percent of patients who will always recommend the hospital. Construct corresponding 95% confidence interval.

   Denote $\tilde{Y}^{80}$ as the predicted percent of patients who always recommend the hospital in a new hospital where patients reporting that nurses always communicate well 80% of the time.

   $\tilde{Y}^{80} = 73.54339$. To find a confidence interval, we need to account for additional uncertainty associate with predicting a new outcome.

   $se(\tilde{Y}^{80}) = \sqrt{var(\bar{Y}^{80}) + \hat{\sigma^2}} = \sqrt{.1399631^2 + 7.2804^2} = 7.281745$

```
. di 73.54339  - invttail(3568, 0.025)*7.281745
59.266589
```

```
. di 73.54339  + invttail(3568, 0.025)*7.281745
87.820191
```

So, a 95% confidence interval $\tilde{Y}^{80}$ is $73.54339 \pm t_{3568,0.975} * 7.281745 = (59.27, 87.82)$.