Multiple Linear Regression

Now, the hospital aims to assess the impact of nurse communication *and* hospital noise level on the percentage of patients who would always recommend the hospital.

Fit a linear regression model with `recommendyes` as the outcome and `nursealways` and `quietalways` as the covariates.

1. Make a scatter plot of `quietalways` versus `recommendyes`.

```
twoway (scatter recommendyes quietalways)
```

While the relationship appears linear, note that we cannot assess any of the assumptions of multiple linear regression using this plot.

2. State your model.

$Yi =$ percent of patients who recommend the hospital always
$X_{1i} =$ percent of patients in a hospital who say that the nurse always communicates well
$X_{2i} =$ percent of patients who report that the hospital is always quiet

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

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

$$\mu_{y_i} = E(Y_i | X_{1i}, X_{2i}) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i}$$

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

3. Fit the model.

```
 regress recommendyes nursealways quietalways

      Source |       SS       df       MS              Number of obs =    3570
-------------+------------------------------           F(  2,  3567) = 1363.40
       Model |  144484.252     2   72242.126           Prob > F      =  0.0000
    Residual |  189003.571  3567  52.9867033           R-squared     =  0.4333
-------------+------------------------------           Adj R-squared =  0.4329
       Total |  333487.823  3569  93.4401297           Root MSE      =  7.2792

------------------------------------------------------------------------------
recommendyes |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
 nursealways |   1.133725   .0282517    40.13   0.000     1.078334    1.189116
 quietalways |   .0229694   .0155642     1.48   0.140    -.0075463     .053485
       _cons |  -18.58225    1.7655   -10.53   0.000    -22.04374   -15.12075
-------------------------------------------------------------------------------
```

Our model is $Y_i = \alpha + 1.13X_{1i} + 0.02X_{2i} + \epsilon_i$, where $\epsilon_i \sim N(0, 7.28^2)$.

4. Evaluate the model assumptions.
   The adjusted R^2 is 0.43 (compared to 0.43 from the simple linear regression model with only `nursealways`).

```
rvfplot
rvpplot nursealways
rvpplot quietalways
```

5. Interpret the coefficients.

   We estimate $\hat{\beta}_1 = 1.13$, with 95% confidence interval $(1.08, 1.19)$. For a one percent increase in the patients who say that the nurses always communicate well, we see on average a 1.13 percent increase in the percent of patients who would always recommend the hospital, when the percent of patients who say the hospital is always quiet is fixed (does not vary).

   We estimate $\hat{\beta}_2 = 0.02$, with 95% confidence interval $(-0.01, 0.05)$. For a one percent increase in the patients who say the hospital is always quiet, we see on average a 0.02 percent increase in the percent of patients who would always recommend the hospital, fixing the percent of patients who say their nurse always communicates well.

   We estimate that $\hat{\alpha} = -18.58$. α is the value of $E(Y_i)$ when X_{1i} and X_{2i} are set to 0. In our dataset, the covariates never drop below 48% and 30% respectively, and therefore α does not have a meaningful interpretation for this study.

6. Suppose we consider a new hospital, where the percentage of nurses who always communicate is 90% and the percentage of those who say the hospital is always quiet is 70%? What is the expected percent of patients who would always recommend this hospital?

   $E(Y_i|X_{1i} = 90, X_{2i} = 70) = -18.58 + 1.13 * 90 + 0.02 * 70 = 84.5\%$.

7. Using the regression results above, perform the follow three hypothesis tests at the 0.05 level of significance.

   - $H_0 : \beta_1 = 0, H_A : \beta_1 \neq 0$

$\hat{\beta}_1 = 1.13$, $\hat{se}(\hat{\beta}_1) = 0.03$, $t = 40.1$. Under $H_0, t \sim t_{3570-2-1}$, and $p < 0.0001$. We reject H_0 and conclude that an increase in the percent of patients who say nurses always communicate well results in an increase in the percent of patients who always recommend the hospital, fixing the percent of patients who say the hospital is always quiet.

- $H_0 : \beta_2 = 0, H_A : \beta_2 \neq 0$

  $\hat{\beta}_1 = 0.02$, $\hat{se}(\hat{\beta}_1) = 0.02$, $t = 1.5$. Under $H_0, t \sim t_{3570-2-1}$, and $p = 0.14$. We fail to reject H_0 and conclude that we do not have evidence in the data that increasing the percent of patients who say the hospital is always quiet is correlated with the percent of patients who always recommend the hospital, fixing the percent of patients who say that the nurses always communicate well .

- $H_0 : \beta_1 = \beta_2 = 0, H_A$: one of$\beta_1, \beta_2 \neq 0$

```
. test nursealways quietalways

 ( 1)  nursealways = 0
 ( 2)  quietalways = 0

       F(  2,  3567) = 1363.40
            Prob > F =    0.0000
```

  Our F-statistic equals 1363.4. Under $H_0, F \sim F_{2,3567}$, and $p < 0.0001$. We reject H_0 and conclude that atleast one of β_1 or β_2 is non-zero.

8. Do we observe any collinearity between X_{1i} and X_{2i}. How does this impact the result.

```
twoway (scatter nursealways quietalways)
```

Yes, the covariates are collinear. We would likely see an association between X_{2i} and Y_i if X_{1i} was excluded from the model.