Indicator Variables and Regression

Suppose a hospital is trying to set a benchmark goal of having patients report that nurses always communicate well at least 75% of the time. We now define a nurse communication indicator variable and use simple linear regression to further examine the relationship between nurse communication and the percentage of patients always recommending the hospital.

Open the dataset `hospitaldata.dta`.

Exercises:

1. Generate a new variable, `highnurse`, that equals 1 if a hospital had `nursealways` $\geq$ 75%; and equals 0 if `nursealways` $<$ 75%.

```
gen highnurse = .
replace highnurse = 1 if nursealways >= 75 & nursealways <= 100
replace highnurse = 0 if nursealways < 75
```

2. State your model and evaluate the model assumptions.

$Y_i =$ percent of patients who recommend the hospital always
$D_i = 1$ if at least 75% of patients at the hospital report that nurses communicate well, and is 0 otherwise

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

where $\epsilon_i \sim N(0, \sigma^2)$.

The model is identical to a one-way ANOVA therefore the assumptions we make are the same. When we only have two groups, the assumptions are identical to the t-test with equal variances.

3. Fit the model.

```
xi: regress recommendyes i.highnurse
```

or

```
regress recommendyes highnurse
      Source |       SS       df       MS              Number of obs =    3570
-------------+------------------------------           F(  1,  3568) = 1004.37
```

```
       Model |  73254.0735     1  73254.0735           Prob > F      =  0.0000
    Residual |  260233.749  3568  72.9354678           R-squared     =  0.2197
-------------+------------------------------           Adj R-squared =  0.2194
       Total |  333487.823  3569  93.4401297           Root MSE      =  8.5402

------------------------------------------------------------------------------
recommendyes |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   highnurse |   9.980834   .3149346    31.69   0.000     9.363364     10.5983
       _cons |   62.86486   .2653319   236.93   0.000     62.34465    63.38508
------------------------------------------------------------------------------
```

So, our fitted model is $Y_i = 62.9 + 10.0 * D_i + \epsilon_i$, where $\epsilon_i \sim N(0, 8.5^2)$.

4. Interpret the coefficients.

   $\hat{\alpha} = 62.9$ is $E(Y_i|D_i = 0)$. The average percent of patients who always recommend a hospital when less than 75% of patients say nurses always communicated well is 62.9%.

   $\hat{\beta} = 10.0$ is $E(Y_i|D_i = 1) - E(Y_i|D_i = 0)$. Comparing hospitals with at least 75% of patients say nurses always communicated well with those where less than 75% of the patients report that nurses always communicate well, the average difference in percent of patients who always recommend a hospital was 10%.

   $\hat{\alpha} + \hat{\beta} = 72.9$ is $E(Y_i|D_i = 1)$. The average percent of patients who always recommend a hospital when at least 75% of patients say nurses always communicated well is 72.9%.

5. Test the null hypothesis that there is no difference in the average percent of patients who always recommend a hospital between hospitals with less than and at least 75% of patients reporting that nurses always communicate well.

   We test $H_0 : \beta = 0$ versus $H_A : \beta \neq 0$ using a two-sided test with $\alpha = 0.05$.
   We find that $\hat{\beta} = 10.0$, $\hat{se}(\hat{\beta}) = 0.3$, and $t = 31.7$. Under H_0, $t \sim t_{n-2}$, and $p < 0.0001$. We conclude that the average percent of patients who always recommend a hospital is greater when at least 75% of patients report that nurses always communicate well.

6. Compare the results of the test above to a two-sample t-test with equal variances.

```
. ttest recommendyes, by(highnurse)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
```

```
---------+--------------------------------------------------------------------
       0 |    1036    62.86486     .272132    8.759099    62.33087    63.39886
       1 |    2534     72.8457    .1678457    8.449162    72.51657    73.17483
---------+--------------------------------------------------------------------
combined |    3570     69.9493    .1617829    9.666443     69.6321     70.2665
---------+--------------------------------------------------------------------
    diff |           -9.980834    .3149346               -10.5983   -9.363364
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -31.6918
Ho: diff = 0                                     degrees of freedom =     3568

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

You should notice some striking similarities!