Lab Assignment 4 BIS 505b

Wenxin Xu

3/28/2021

Contents

Instructions

This Lab Assignment uses the data from the study conducted to investigate the impacts of herbicide exposure on maternal health described in **Lab Assignment 0**, `hgb.csv`. We would like to explore the association between tap water consumption group and hemoglobin change (g/dL), controlling for potentially important variables or confounders of the association using multiple linear regression. In this assignment, report any p-values that are less than 0.001 as **<0.001** and round values reported in your narrative text to **3** decimal places. **Be sure to clearly state the reference category when interpreting the effects of categorical variables in any regression model.**

Assignment

1. [5 points] Import the CSV file `hgb.csv` in the third code chunk above. Name your data frame `hgb` and re-create the variables `group_factor` (reference = Bottled only (code provided below)), `prenatal_factor` (reference = No) and `psmoke_factor` (reference = No) that you created in **Lab Assignment 0**.

Instead of analyzing `change` in this lab, we will analyze `hgbdecline = hgb$hgb9 - hgb$hgb36`, or week 9 hemoglobin [`hgb9`] minus the week 36 hemoglobin [`hgb36`]. Since hemoglobin decreases during pregnancy for all women in our data set, the variable `hgbdecline` will be positive for all individuals. A larger value for `hgbdecline` indicates that hemoglobin decreased a greater amount during pregnancy. When modeling `hgbdecline` as the response, positive slopes indicate that hemoglobin is declining more (greater decline from baseline), while negative slopes indicate that hemoglobin is declining less (smaller decline from baseline).

After these steps, `hgb` should contain 17 variables. [**Note:** When creating factor variables, **do not** use the `ordered=TRUE` option to create ordinal variables. No written response is required for this question. Display the code chunk(s) that perform the requested data management steps for this question.]

```
hgb$hgbdecline <- hgb$hgb9 - hgb$hgb36

hgb$group_factor <- factor(hgb$group,
                           levels = c(2, 3, 1),
                           labels = c("Bottled only",
```

```
                                        "Combination",
                                        "Tap only"))
hgb$prenatal_factor <- factor(hgb$prenatal,
                              levels = c(0,1),
                              labels=c("No","Yes"))

hgb$psmoke_factor <- factor(hgb$psmoke,
                            levels=c(0,1),
                           labels = c("No","Yes"))

# now data frame hgb has 17 variables
ncol(hgb)
```

```
## [1] 17
```

2. The **research question** is: Is type of water consumed [group_factor] associated with hemoglobin decline during pregnancy [hgbdecline]?

a. [20 points] Use water consumption group [group_factor] to model hemoglobin decline [hgbdecline] using a linear regression model.

- How many dummy variables represent group_factor in this model? What is the reference category of group_factor?

```
contrasts(hgb$group_factor)
```

```
##              Combination Tap only
## Bottled only           0        0
## Combination            1        0
## Tap only               0        1
```

group_factor has 2 dummy variables. The reference category is Bottled only group.

- Write the fitted model and interpret the regression parameters (intercept and two slopes). Do the slopes indicate that hemoglobin is declining more (greater decline from baseline) or less (smaller decline from baseline) in the groups being compared?

```
mod.group <- lm(hgbdecline ~ group_factor, data=hgb)
summary(mod.group)
```

```
##
## Call:
## lm(formula = hgbdecline ~ group_factor, data = hgb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.27477 -0.27900 -0.00477  0.28740  1.41740
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2.03260    0.02394   84.89   <2e-16
```

```
## group_factorCombination  1.30217    0.03212   40.54   <2e-16
## group_factorTap only     1.43640    0.03524   40.76   <2e-16
##
## Residual standard error: 0.4249 on 976 degrees of freedom
## Multiple R-squared:  0.6921, Adjusted R-squared:  0.6915
## F-statistic:  1097 on 2 and 976 DF,  p-value: < 2.2e-16
```

```
confint(mod.group)
```

```
##                              2.5 %   97.5 %
## (Intercept)              1.985618 2.079588
## group_factorCombination 1.239141 1.365196
## group_factorTap only    1.367237 1.505557
```

The fitted model is: $\hat{y} = 2.033 + 1.302$ Combination $+ 1.436$ Tap only.

The average hemoglobin decline in Bottled only group is $a = 2.033$ [95% CI (1.986,2.08)]

The average hemoglobin decline in Combination group is $b_1 = 1.302$ [95% CI (1.239, 1.365)] units higher than that in Bottled only group.

The average hemoglobin decline in Tap only group is $b_2 = 1.436$ [95% CI (1.367, 1.506)] units higher than that in Bottled only group.

- Perform three hypothesis tests each at the $\alpha = 0.05$-level to test the following null hypotheses: (1) $H_0 : \beta_1 = 0$, (2) $H_0 : \beta_2 = 0$, and (3) $H_0 : \beta_1 = \beta_2 = 0$. For each test, (i) State the null and alternative hypotheses; (ii) From your **R** output, report the value of the test statistic and p-value; (iii) State your statistical conclusion and your conclusion in the context of the problem.

(1) $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

T test statistic is 40.544, p value is <.001. We reject H_0 and conclude that there is a significant difference in the average hemoglobin decline in Combination group vs. Bottled only group.

(2) $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 \neq 0$

T test statistic is 40.758, p-value <.001. We reject H_0 and conclude that there is a significant difference in the average hemoglobin decline in Combination group vs. Bottled only group.

(3) $H_0 : \beta_1 = \beta_2 = 0$ vs. $H_1 : \beta_1, \beta_2$ not all 0.

The F-statistic is 1096.957, p-value <.001. We reject H_0 and conclude that there is at least one significant difference in the average hemoglobin decline in Combination group vs. Bottled only group or Tap only group vs. Bottled only group.

b. [30 points] Perhaps the effect we observe due to water consumption group is driven by demographic characteristics. For example, maybe women who drink only tap water were more likely to be pre-pregnancy smokers or are less likely to receive adequate prenatal care, and these factors are driving the larger decline in hemoglobin in this group. Our goal is to control for additional characteristics of the mother and examine the *adjusted* effect of type of water consumed [`group_factor`]. Build a multiple linear regression model that controls for income [`income`], number of previous births [`parity`], adequate prenatal care [`prenatal_factor`], and pre-pregnancy smoking status [`psmoke_factor`].

```r
mod.mlr <- lm(hgbdecline ~ group_factor + income + parity + prenatal_factor + psmoke_factor, data= hgb)
summary(mod.mlr)
```

```
##
## Call:
## lm(formula = hgbdecline ~ group_factor + income + parity + prenatal_factor +
##     psmoke_factor, data = hgb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30255 -0.21645 -0.00173  0.22483  1.07691
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.083328   0.043446  47.952  < 2e-16
## group_factorCombination   1.197116   0.026529  45.125  < 2e-16
## group_factorTap only      1.237535   0.030804  40.175  < 2e-16
## income                   -0.025701   0.006572  -3.911 9.85e-05
## parity                    0.234239   0.014984  15.633  < 2e-16
## prenatal_factorYes       -0.307033   0.023927 -12.832  < 2e-16
## psmoke_factorYes          0.224742   0.025532   8.802  < 2e-16
##
## Residual standard error: 0.3388 on 968 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8011, Adjusted R-squared:  0.7999
## F-statistic: 649.8 on 6 and 968 DF,  p-value: < 2.2e-16
```

```r
confint(mod.mlr)
```

```
##                               2.5 %      97.5 %
## (Intercept)               1.9980681  2.16858728
## group_factorCombination   1.1450558  1.24917708
## group_factorTap only      1.1770852  1.29798507
## income                   -0.0385981 -0.01280383
## parity                    0.2048352  0.26364337
## prenatal_factorYes       -0.3539873 -0.26007933
## psmoke_factorYes          0.1746369  0.27484724
```

- Write the fitted model and interpret the regression parameters (slopes). Based on the p-values reported in the **R** output for each slope, state your conclusion about each effect in the MLR model in the context of the problem.

The fitted model is $\hat{y} = 2.083 + 1.197$ Combination $+ 1.238$ Tap only $- 0.026$ Income $+ 0.234$ Parity $- 0.307$ Prenatal $+ 0.225$ Psmoke.

Income, number of previous births, adequate prenatal care and pre-pregnancy smoking status-adjusted effect of water consumption on hemoglobin decline:

The estimated slope of b_1 indicates that the average hemoglobin decline in Combination group is 1.197 [95% CI (1.145, 1.249)] units higher than that in Bottled only group, controlling for income, number of previous births, adequate prenatal care and pre-pregnancy smoking status.

A **significance test** of β_1 shows that there is a significant difference in the average hemoglobin decline in Combination group vs. Bottled only group (p-value $<.001$) when controlling for income, number of previous births, adequate prenatal care and pre-pregnancy smoking status.

The estimated slope of b_2 indicates that the average hemoglobin decline in Tap only group is 1.238 [95% CI (1.177, 1.298)] units higher than that in Bottled only group, controlling for income, number of previous births, adequate prenatal care and pre-pregnancy smoking status.

A **significance test** of β_2 shows that there is a significant difference in the average hemoglobin decline in Tap only group vs. Bottled only group (p-value <.001) when controlling for income, number of previous births, adequate prenatal care and pre-pregnancy smoking status.

- Comment on how the adjusted effects of `group_factor` on `hgbdecline` have changed in this model (adjusted model) compared to the model you constructed in question **2a** (unadjusted model).

The adjusted effects of `group_factor` on `hgbdecline` are smaller than the unadjusted effects. b_1: 1.197 (adjusted) vs.1.302 (unadjusted), b_2: 1.238 (adjusted) vs.1.436 (unadjusted).

- Report the adjusted R^2 of this model to 4 decimal places.

The adjusted R^2 is 0.7999.

c. [10 points] Modify the adjusted model from question **2b** to additionally control for baseline hemoglobin `hgb9`.

```
mod.mlr2 <- lm(hgbdecline ~ group_factor + income + parity + prenatal_factor + psmoke_factor+hgb9, data=
summary(mod.mlr2)
```

```
##
## Call:
## lm(formula = hgbdecline ~ group_factor + income + parity + prenatal_factor +
##     psmoke_factor + hgb9, data = hgb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30896 -0.22082 -0.00518  0.22659  1.08346
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.002283   0.185822  10.775  < 2e-16
## group_factorCombination   1.198124   0.026635  44.984  < 2e-16
## group_factorTap only      1.242593   0.032814  37.868  < 2e-16
## income                   -0.025661   0.006575  -3.903 0.000102
## parity                    0.234146   0.014991  15.619  < 2e-16
## prenatal_factorYes       -0.306638   0.023953 -12.802  < 2e-16
## psmoke_factorYes          0.225122   0.025557   8.809  < 2e-16
## hgb9                      0.007055   0.015727   0.449 0.653833
##
## Residual standard error: 0.339 on 967 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.8012, Adjusted R-squared:  0.7997
## F-statistic: 556.6 on 7 and 967 DF,  p-value: < 2.2e-16
```

- Is the effect of baseline hemoglobin statistically significant in the presence of all of these other predictors?

(i) State the null and alternative hypotheses; $H_0 : \beta_7 = 0$ vs. $H_1 : \beta_7 \neq 0$

(ii) From your **R** output, report the value of the test statistic and p-value.

The t test statistic is 0.449, the p-value = 0.654.

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We fail to reject H_0 and conclude that there is no significant association between baseline hemoglobin and average hemoglobin decline, controlling for water consumption, income, number of previous births, adequate prenatal care and pre-pregnancy smoking status.

- Report the adjusted R^2 of this model to 4 decimal places. Has the adjusted R^2 improved compared to the model in question **2b**? Is `hgb9` adding to the predictive ability of the model?

The adjusted R^2 is 0.7997. The adjusted R^2 decreased, so it did not improved compared to the model in question **2b**. Thus, `hgb9` is not adding to the predictive ability of the model.

d. [35 points] Expand on the model fit in question **2a** to include the main effect of income and an interaction term between water consumption group and income.

```
# interaction model of water consumption and income
mod.intx <- lm(hgbdecline ~ income + group_factor + income*group_factor,
               data=hgb)
summary(mod.intx)
```

```
##
## Call:
## lm(formula = hgbdecline ~ income + group_factor + income * group_factor,
##     data = hgb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41533 -0.28416 -0.00225  0.28152  1.38871
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.095360   0.058266  35.962  < 2e-16
## income                         -0.013313   0.013154  -1.012  0.31173
## group_factorCombination         1.274589   0.078229  16.293  < 2e-16
## group_factorTap only            1.691272   0.091324  18.519  < 2e-16
## income:group_factorCombination  0.005813   0.017954   0.324  0.74617
## income:group_factorTap only    -0.071309   0.021896  -3.257  0.00117
##
## Residual standard error: 0.4118 on 969 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7059, Adjusted R-squared:  0.7043
## F-statistic: 465.1 on 5 and 969 DF,  p-value: < 2.2e-16
```

- Write the fitted model.

The fitted model is $\hat{y} = 2.095 - 0.013$ Income $+ 1.275$ Combination $+ 1.691$ Tap only $+ 0.006$ Income $\times$ Combination $- 0.071$ Income $\times$ Tap only.

- Write the model $\mu_{y|x}$ in each of the three water consumption groups (bottled only group, combination group, and tap only group).

Model for bottled only group:

$\mu_{y|x_1,z_1=0,z_2=0} = \alpha + \beta_1 x_1 + \beta_2(0) + \beta_3(0) + \beta_4 x_1(0) + \beta_5 x_1(0) = \alpha + \beta_1 x_1$

Model for combination group:

$\mu_{y|x_1,z_1=1,z_2=0} = \alpha + \beta_1 x_1 + \beta_2 + \beta_3(0) + \beta_4 x_1 + \beta_5 x_1(0) = \alpha + \beta_2 + (\beta_1 + \beta_4)x_1$

Model for tap only group:

$\mu_{y|x_1,z_1=0,z_2=1} = \alpha + \beta_1 x_1 + \beta_2(0) + \beta_3 + \beta_4 x_1(0) + \beta_5 x_1 = \alpha + \beta_3 + (\beta_1 + \beta_5)x_1$

- Perform a partial F-test to simultaneously test the model parameters (βs) involved in the interaction. (i) State the null and alternative hypotheses; (ii) From your **R** output, report the value of the test statistic and p-value; (iii) State your statistical conclusion and your conclusion in the context of the problem.

```
# full model is the previous interaction model

# reduced model, under H_0, don't include interaction term
mod.red <- lm(hgbdecline ~ income + group_factor,
              data=mod.intx$model)

# F-test comparing full and reduced models
anova(mod.red, mod.intx)
```

```
## Analysis of Variance Table
##
## Model 1: hgbdecline ~ income + group_factor
## Model 2: hgbdecline ~ income + group_factor + income * group_factor
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    971 166.80
## 2    969 164.35  2    2.4485 7.2179 0.0007735
```

(i) State the null and alternative hypotheses;

$H_0 : \beta_4 = \beta_5 = 0$ vs. $H_1 : \beta_4, \beta_5$ not all 0.

(ii) From your **R** output, report the value of the test statistic and p-value;

The F test statistic is 7.218, the p-value <.001.

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We reject H_0 and conclude that there is at least one significant difference in the effect of income on hemoglobin decline in the combination group vs. bottled only group or in the tap only group vs. bottled only group.

- Tease apart the interaction to report and interpret the effect of income in each water consumption group (bottled only group, combination group, and tap only group). Perform a hypothesis test to determine if the effect of income is statistically significant in each group.

1)The effect of income on hemoglobin decline in the bottled only group is estimated by b_1

The tease apart part is wrong. for instance, for combination group, you should use rbind(c(0, 0, 0, 1, 1, 0)) , and test b3+b4=0

```r
# b1: effect of income in the bottled only group

# the same interaction model used in question d
mod.intx <- lm(hgbdecline ~ income + group_factor + income*group_factor,
               data=hgb)
summary(mod.intx)
```

```
##
## Call:
## lm(formula = hgbdecline ~ income + group_factor + income * group_factor,
##     data = hgb)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.41533 -0.28416 -0.00225  0.28152  1.38871
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.095360   0.058266  35.962  < 2e-16
## income                         -0.013313   0.013154  -1.012  0.31173
## group_factorCombination         1.274589   0.078229  16.293  < 2e-16
## group_factorTap only            1.691272   0.091324  18.519  < 2e-16
## income:group_factorCombination  0.005813   0.017954   0.324  0.74617
## income:group_factorTap only    -0.071309   0.021896  -3.257  0.00117
##
## Residual standard error: 0.4118 on 969 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.7059, Adjusted R-squared:  0.7043
## F-statistic: 465.1 on 5 and 969 DF,  p-value: < 2.2e-16
```

(i) State the null and alternative hypotheses;

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

(ii) From your **R** output, report the p-value of this test;

The p-value = 0.312

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We fail to reject H_0 and conclude that there is no significant association between income and hemoglobin decline in the bottled only group.

2) The effect of income on hemoglobin decline in the combination group is estimated by $b_1 + b_4$.

```r
# b1 + b4: effect of income in the combination group

# vector that specifies linear combination of coefficients interested in
K1 <- rbind(c(0,1,0,0,1,0))

# label for comparison (printed in the output)
```

```r
rownames(K1) <- "b1+b4 (slope in group_factor=combination)"

# estimate of slope (b1+b4) and hypothesis test
summary(glht(mod.intx,linfct=K1))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = hgbdecline ~ income + group_factor + income * group_factor,
##     data = hgb)
##
## Linear Hypotheses:
##                                                 Estimate Std. Error t value
## b1+b4 (slope in group_factor=combination) == 0 -0.00750    0.01222  -0.614
##                                                 Pr(>|t|)
## b1+b4 (slope in group_factor=combination) == 0     0.54
## (Adjusted p values reported -- single-step method)
```

```r
# confidence interval for beta1+beta3
confint(glht(mod.intx,linfct=K1))
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = hgbdecline ~ income + group_factor + income * group_factor,
##     data = hgb)
##
## Quantile = 1.9624
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                                                 Estimate lwr      upr
## b1+b4 (slope in group_factor=combination) == 0 -0.00750 -0.03148  0.01648
```

(i) State the null and alternative hypotheses;

$H_0 : \beta_1 + \beta_4 = 0$ vs. $H_1 : \beta_1 + \beta_4 \neq 0$

(ii) From your **R** output, report the p-value of this test;

The p-value = 0.54.

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We fail to reject H_0 and conclude that there is no significant association between income and hemoglobin decline in the combination group.

3) The effect of income on hemoglobin decline in the tap only group is estimated by $b_1 + b_5$.

```r
# b1 + b5: effect of income in the tap only group

# vector that specifies linear combination of coefficients interested in
K2 <- rbind(c(0,1,0,0,0,1))

# label for comparison (printed in the output)
rownames(K2) <- "b1+b4 (slope in group_factor=tap only)"

# estimate of slope (b1+b5) and hypothesis test
summary(glht(mod.intx,linfct=K2))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = hgbdecline ~ income + group_factor + income * group_factor,
##     data = hgb)
##
## Linear Hypotheses:
##                                            Estimate Std. Error t value
## b1+b4 (slope in group_factor=tap only) == 0 -0.08462    0.01750  -4.834
##                                            Pr(>|t|)
## b1+b4 (slope in group_factor=tap only) == 0 1.55e-06
## (Adjusted p values reported -- single-step method)
```

```r
confint(glht(mod.intx,linfct=K2))
```

```
##
##   Simultaneous Confidence Intervals
##
## Fit: lm(formula = hgbdecline ~ income + group_factor + income * group_factor,
##     data = hgb)
##
## Quantile = 1.9624
## 95% family-wise confidence level
##
##
## Linear Hypotheses:
##                                            Estimate lwr      upr
## b1+b4 (slope in group_factor=tap only) == 0 -0.08462 -0.11897 -0.05027
```

(i) State the null and alternative hypotheses;

$H_0 : \beta_1 + \beta_5 = 0$ vs. $H_1 : \beta_1 + \beta_5 \neq 0$

(ii) From your **R** output, report the p-value of this test;

The p-value $<.001$.

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We reject H_0 and conclude that there is significant association between income and hemoglobin decline in the tap only group.