Assignment3

jessekim

10/26/2019

2013-11086 Chankyu Kim

Assignment 3

---

1. EE3.1

```r
head(cps)
```

```
## # A tibble: 6 x 5
##    year   ahe bachelor female   age
##   <dbl> <dbl>    <dbl>  <dbl> <dbl>
## 1  1996 11.2         0      0    31
## 2  1996  8.65        0      1    31
## 3  1996  9.62        1      1    27
## 4  1996 11.2         1      0    26
## 5  1996  9.62        1      1    28
## 6  1996 14.4         1      0    32
```

a.

```r
(cps_params <- cps %>%
  group_by(year) %>%
  dplyr::summarize(
    n = n(),
    mean_ahe = mean(ahe),
    sd_ahe = sd(ahe),
    se_ahe = sd_ahe/sqrt(n),
    inf_ci = mean_ahe - 1.96*se_ahe,
    sup_ci = mean_ahe + 1.96*se_ahe
  ))
```

```
## # A tibble: 2 x 7
##    year     n mean_ahe sd_ahe se_ahe inf_ci sup_ci
##   <dbl> <int>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
## 1  1996  6103     12.7   6.36 0.0814   12.5   12.9
## 2  2015  7098     21.2   12.1 0.144    21.0   21.5
```

i.

Sample mean for AHE in 1996: 12.69. Sample mean for AHE in 2015: 21.24.

ii.

Sample standard deviation for AHE in 1996: 6.36. Sample standard deviation for AHE in 2015: 12.12.

iii.

CI for the pop mean of AHE in 1996: (12.53, 12.85). CI for the pop mean of AHE in 2015: (20.96, 21.52).

iv.

```
tibble(
  mean_diff = cps_params[["mean_ahe"]][2] - cps_params[["mean_ahe"]][1],
  se_diff = sqrt(cps_params[["se_ahe"]][2]^2 + cps_params[["se_ahe"]][1]^2),
  inf_ci = mean_diff - 1.96*se_diff,
  sup_ci = mean_diff + 1.96*se_diff
)
```

```
## # A tibble: 1 x 4
##   mean_diff se_diff inf_ci sup_ci
##       <dbl>   <dbl>  <dbl>  <dbl>
## 1      8.54   0.165   8.22   8.87
```

CI for the change in the pop mean of AHE: (8.22, 8.87).

b.

Adjustment:

```
cpi <- tibble(
  year = c(1996, 2015),
  cpi = c(156.9, 237.0)
)

cps_adj <- cps %>%
  inner_join(cpi, by = c("year")) %>%
  mutate(ahe_adj = ahe*237/cpi)

head(cps_adj)
```

```
## # A tibble: 6 x 7
##    year   ahe bachelor female   age   cpi ahe_adj
##   <dbl> <dbl>    <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1  1996 11.2         0      0    31  157.    16.9
## 2  1996  8.65        0      1    31  157.    13.1
## 3  1996  9.62        1      1    27  157.    14.5
## 4  1996 11.2         1      0    26  157.    16.9
## 5  1996  9.62        1      1    28  157.    14.5
## 6  1996 14.4         1      0    32  157.    21.8
```

```
(cps_adj_params <- cps_adj %>%
  group_by(year) %>%
  dplyr::summarize(
    n = n(),
    mean_ahe = mean(ahe_adj),
    sd_ahe = sd(ahe_adj),
    se_ahe = sd_ahe/sqrt(n),
    inf_ci = mean_ahe - 1.96*se_ahe,
    sup_ci = mean_ahe + 1.96*se_ahe
  ))
```

```
## # A tibble: 2 x 7
##    year     n mean_ahe sd_ahe se_ahe inf_ci sup_ci
##   <dbl> <int>    <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
```

```
## 1  1996  6103     19.2   9.61  0.123   18.9   19.4
## 2  2015  7098     21.2  12.1   0.144   21.0   21.5
```

i.

Sample mean for adjusted AHE in 1996: 19.17. Sample mean for adjusted AHE in 2015: 21.24.

ii.

Sample standard deviation for adjusted AHE in 1996: 9.61. Sample standard deviation for adjusted AHE in 2015: 12.12.

iii.

CI for the pop mean of adjusted AHE in 1996: (18.93, 19.41). CI for the pop mean of adjusted AHE in 2015: (20.96, 21.52).

iv.

```
tibble(
  mean_diff = cps_adj_params[["mean_ahe"]][2] - cps_adj_params[["mean_ahe"]][1],
  se_diff = sqrt(cps_adj_params[["se_ahe"]][2]^2 + cps_adj_params[["se_ahe"]][1]^2),
  inf_ci = mean_diff - 1.96*se_diff,
  sup_ci = mean_diff + 1.96*se_diff
)
```

```
## # A tibble: 1 x 4
##   mean_diff se_diff inf_ci sup_ci
##       <dbl>   <dbl>  <dbl>  <dbl>
## 1      2.06   0.189   1.69   2.44
```

CI for the change in the pop mean of adjusted AHE: (1.69, 2.44).

c.

Note that the prices of the commodities change year by year. Thus the results from (b), adjusted by CPI(Customer Price Index), reflects real purchasing power of the commodities better.

d.

```
(ahe_2015_bch <- cps_adj %>%
  filter(year == 2015) %>%
  group_by(bachelor) %>%
  dplyr::summarize(
    n = n(),
    mean_ahe = mean(ahe_adj),
    se_ahe = sd(ahe_adj)/sqrt(n),
    inf_ci = mean_ahe - 1.96*se_ahe,
    sup_ci = mean_ahe + 1.96*se_ahe
  ))
```

```
## # A tibble: 2 x 6
##   bachelor     n mean_ahe se_ahe inf_ci sup_ci
##      <dbl> <int>    <dbl>  <dbl>  <dbl>  <dbl>
## 1        0  3365     16.4  0.147   16.1   16.7
## 2        1  3733     25.6  0.216   25.2   26.0
```

i.

CI for the mean of AHE for high schoolers: (16.09, 16.67).

ii.

CI for the mean of AHE for college graudates: (25.19, 26.04).

iii.

```
tibble(
  mean_diff = ahe_2015_bch[["mean_ahe"]][2] - ahe_2015_bch[["mean_ahe"]][1],
  se_diff = sqrt(ahe_2015_bch[["se_ahe"]][2]^2 + ahe_2015_bch[["se_ahe"]][1]^2),
  inf_ci = mean_diff - 1.96*se_diff,
  sup_ci = mean_diff + 1.96*se_diff
)
```

```
## # A tibble: 1 x 4
##   mean_diff se_diff inf_ci sup_ci
##       <dbl>   <dbl>  <dbl>  <dbl>
## 1      9.23   0.261   8.72   9.75
```

CI for the mean difference of AHE: (8.72, 9.75).

e.

```
(ahe_1996_bch <- cps_adj %>%
  filter(year == 1996) %>%
  group_by(bachelor) %>%
  dplyr::summarize(
    n = n(),
    mean_ahe = mean(ahe_adj),
    se_ahe = sd(ahe_adj)/sqrt(n),
    inf_ci = mean_ahe - 1.96*se_ahe,
    sup_ci = mean_ahe + 1.96*se_ahe
  ))
```

```
## # A tibble: 2 x 6
##   bachelor     n mean_ahe se_ahe inf_ci sup_ci
##      <dbl> <int>    <dbl>  <dbl>  <dbl>  <dbl>
## 1        0  3484     16.3  0.130   16.0   16.5
## 2        1  2619     23.0  0.205   22.6   23.4
```

i.

CI for the mean of AHE for high schoolers: (16.01, 16.52).

ii.

CI for the mean of AHE for college graudates: (22.64, 23.44).

iv.

```
tibble(
  mean_diff = ahe_1996_bch[["mean_ahe"]][2] - ahe_1996_bch[["mean_ahe"]][1],
  se_diff = sqrt(ahe_1996_bch[["se_ahe"]][2]^2 + ahe_1996_bch[["se_ahe"]][1]^2),
  inf_ci = mean_diff - 1.96*se_diff,
  sup_ci = mean_diff + 1.96*se_diff
)
```

```
## # A tibble: 1 x 4
##   mean_diff se_diff inf_ci sup_ci
##       <dbl>   <dbl>  <dbl>  <dbl>
## 1      6.77   0.243   6.29   7.25
```

CI for the mean difference of AHE: (6.29, 7.25).

f.

```r
(cps_year_bach <- cps_adj %>%
  group_by(year, bachelor) %>%
  dplyr::summarize(
    n = n(),
    mean_ahe = mean(ahe_adj),
    se_ahe = sd(ahe_adj)/sqrt(n)
  ))
```

```
## # A tibble: 4 x 5
## # Groups:   year [2]
##    year bachelor     n mean_ahe se_ahe
##   <dbl>    <dbl> <int>    <dbl>  <dbl>
## 1  1996        0  3484     16.3  0.130
## 2  1996        1  2619     23.0  0.205
## 3  2015        0  3365     16.4  0.147
## 4  2015        1  3733     25.6  0.216
```

i.

```r
mean_chg_high <- cps_year_bach[["mean_ahe"]][3] - cps_year_bach[["mean_ahe"]][1]
se_chg_high <- sqrt(cps_year_bach[["se_ahe"]][3]^2 + cps_year_bach[["se_ahe"]][1]^2)
t_chg_high <- mean_chg_high/se_chg_high
sprintf("T statistic for change in ahe for high school graduates is %.4f", t_chg_high)
```

```
## [1] "T statistic for change in ahe for high school graduates is 0.5749"
```

So there is no statistically significance evidence that AHE for high school graduates increased from 1996 to 2015.

```r
mean_chg_coll <- cps_year_bach[["mean_ahe"]][4] - cps_year_bach[["mean_ahe"]][2]
se_chg_coll <- sqrt(cps_year_bach[["se_ahe"]][4]^2 + cps_year_bach[["se_ahe"]][2]^2)
t_chg_coll <- mean_chg_coll/se_chg_coll
sprintf("T statistic for change in ahe for college graduates is %.4f", t_chg_coll)
```

```
## [1] "T statistic for change in ahe for college graduates is 8.6540"
```

So there is statistically significance evidence that AHE for college graduates increased from 1996 to 2015.

```r
mean_chg_diff <- mean_chg_coll - mean_chg_high
se_chg_diff <- sqrt(se_chg_coll^2 + se_chg_high^2)
t_chg_diff <- mean_chg_diff/se_chg_diff
sprintf("T statistic for change in ahe gap is %.4f", t_chg_diff)
```

```
## [1] "T statistic for change in ahe gap is 6.9084"
```

So there is statistically significance evidence that AHE gap increased from 1996 to 2015.

g.

Gender gap for high school graudates

```r
gender_gap <- cps_adj %>%
  filter(bachelor == 0) %>%
  group_by(year, female) %>%
```

```r
  dplyr::summarise(
    mean = mean(ahe_adj),
    sd = sd(ahe_adj),
    n = n()
  )

mean_diff_1996 <- gender_gap[["mean"]][1] - gender_gap[["mean"]][2]
mean_diff_2015 <- gender_gap[["mean"]][3] - gender_gap[["mean"]][4]

se_diff_1996 <- sqrt(
  gender_gap[["sd"]][1]^2/gender_gap[["n"]][1] + gender_gap[["sd"]][2]^2/gender_gap[["n"]][2]
)
se_diff_2015 <- sqrt(
  gender_gap[["sd"]][3]^2/gender_gap[["n"]][3] + gender_gap[["sd"]][4]^2/gender_gap[["n"]][4]
)

diff <- tibble(
  year = c(1996, 2015),
  mean_diff = c(mean_diff_1996, mean_diff_2015),
  se_diff = c(se_diff_1996, se_diff_2015),
  CI_inf = c(mean_diff_1996 - 1.96*se_diff_1996, mean_diff_2015 - 1.96*se_diff_2015),
  CI_sup = c(mean_diff_1996 + 1.96*se_diff_1996, mean_diff_2015 + 1.96*se_diff_2015)
)

gender_gap %>%
  filter(female == 0) %>%
  left_join(
    gender_gap %>%
      filter(female == 1),
    by = c("year"),
    suffix = c("(Men)", "(Women)")
  ) %>%
  left_join(
    diff,
    by = "year"
  ) %>%
  dplyr::select(-c("female(Men)", "female(Women)"))
```

```
## # A tibble: 2 x 11
## # Groups:   year [2]
##    year `mean(Men)` `sd(Men)` `n(Men)` `mean(Women)` `sd(Women)` `n(Women)`
##   <dbl>       <dbl>     <dbl>    <int>         <dbl>       <dbl>      <int>
## 1  1996        17.8      8.24     2168          13.8        5.83       1316
## 2  2015        17.5      9.03     2222          14.2        7.00       1143
## # ... with 4 more variables: mean_diff <dbl>, se_diff <dbl>, CI_inf <dbl>,
## #   CI_sup <dbl>
```

Gender gap for college graudates

```r
gender_gap_coll <- cps_adj %>%
  filter(bachelor == 1) %>%
  group_by(year, female) %>%
  dplyr::summarise(
    mean = mean(ahe_adj),
```

```r
    sd = sd(ahe_adj),
    n = n()
  )

mean_diff_1996_coll <- gender_gap_coll[["mean"]][1] - gender_gap_coll[["mean"]][2]
mean_diff_2015_coll <- gender_gap_coll[["mean"]][3] - gender_gap_coll[["mean"]][4]

se_diff_1996_coll <- sqrt(
  gender_gap_coll[["sd"]][1]^2/gender_gap_coll[["n"]][1] + gender_gap_coll[["sd"]][2]^2/gender_gap_coll
)
se_diff_2015_coll <- sqrt(
  gender_gap_coll[["sd"]][3]^2/gender_gap_coll[["n"]][3] + gender_gap_coll[["sd"]][4]^2/gender_gap_coll
)

diff_coll <- tibble(
  year = c(1996, 2015),
  mean_diff = c(mean_diff_1996_coll, mean_diff_2015_coll),
  se_diff = c(se_diff_1996_coll, se_diff_2015_coll),
  CI_inf = c(mean_diff_1996_coll - 1.96*se_diff_1996_coll, mean_diff_2015_coll - 1.96*se_diff_2015_coll
  CI_sup = c(mean_diff_1996_coll + 1.96*se_diff_1996_coll, mean_diff_2015_coll + 1.96*se_diff_2015_coll
)

gender_gap_coll %>%
  filter(female == 0) %>%
  left_join(
    gender_gap_coll %>%
      filter(female == 1),
    by = c("year"),
    suffix = c("(Men)", "(Women)")
  ) %>%
  left_join(
    diff_coll,
    by = "year"
  ) %>%
  dplyr::select(-c("female(Men)", "female(Women)"))
```

```
## # A tibble: 2 x 11
## # Groups:   year [2]
##    year `mean(Men)` `sd(Men)` `n(Men)` `mean(Women)` `sd(Women)` `n(Women)`
##   <dbl>       <dbl>     <dbl>    <int>         <dbl>       <dbl>      <int>
## 1  1996        24.9      11.4     1387          21.0        8.93       1232
## 2  2015        28.1      14.4     1917          23.0       11.2        1816
## # ... with 4 more variables: mean_diff <dbl>, se_diff <dbl>, CI_inf <dbl>,
## #   CI_sup <dbl>
```

As the gender gap for college graduates increased from 1996 (3.88) to 2015 (5.02), the gender gap for high school graduates decreased from 1996 (4.02) to 2015 (3.29). For all periods, the gender gaps for college/high school graduates are both statistically significant.

---

2. EE4.2

```r
head(eah)
```

```
## # A tibble: 6 x 11
##     sex   age   mrd  educ cworker region  race earnings height weight
##   <dbl> <dbl> <dbl> <dbl>   <dbl>  <dbl> <dbl>    <dbl>  <dbl>  <dbl>
## 1     0    48     1    13       1      3     1   84055.     65    133
## 2     0    41     6    12       1      2     1   14021.     65    155
## 3     0    26     1    16       1      1     1   84055.     60    108
## 4     0    37     1    16       1      2     1   84055.     67    150
## 5     0    35     6    16       1      1     1   28560.     68    180
## 6     0    25     6    15       1      4     1   23363.     63    101
## # ... with 1 more variable: occupation <dbl>
```

a.

```
eah[["height"]] %>% median()
```

```
## [1] 67
```

The median value is 67.

b.

```
(eah_params <- eah %>%
  mutate(is_tall = height > 67) %>%
  group_by(is_tall) %>%
  dplyr::summarize(
    n = n(),
    mean_earnings = mean(earnings),
    se_earnings = sd(earnings)/sqrt(n)
  ))
```

```
## # A tibble: 2 x 4
##   is_tall     n mean_earnings se_earnings
##   <lgl>   <int>         <dbl>       <dbl>
## 1 FALSE   10114        44488.        265.
## 2 TRUE     7756        49988.        305.
```

i.

The sample mean of earnings for workers who are not tall: 44,488.44.

ii.

The sample mean of earnings for workers who are tall: 49,987.88.

iii.

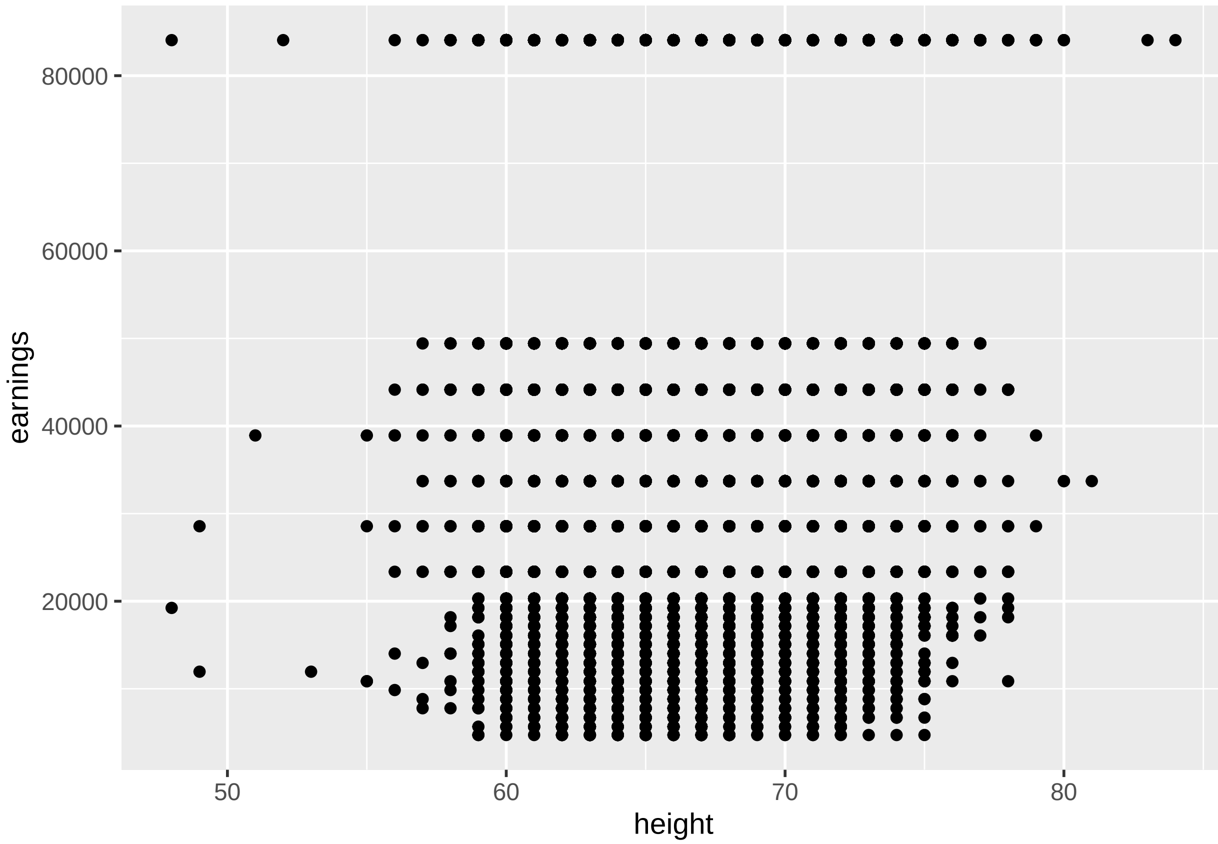
```
tibble(
  mean_diff = eah_params[["mean_earnings"]][2] - eah_params[["mean_earnings"]][1],
  se_diff = sqrt(eah_params[["se_earnings"]][2]^2 + eah_params[["se_earnings"]][1]^2),
  inf_ci = mean_diff - 1.96*se_diff,
  sup_ci = mean_diff + 1.96*se_diff
)
```

```
## # A tibble: 1 x 4
##   mean_diff se_diff inf_ci sup_ci
##       <dbl>   <dbl>  <dbl>  <dbl>
## 1     5499.    405.  4706.  6293.
```

The mean difference between the workers who are tall and not tall is 5,499.44. The taller workers earn more in average, by 5,499.44. The CI for the difference is (4,706.28, 6,292.60)

c.

```r
ggplot(eah, aes(height, earnings)) +
  geom_point()
```



Data description says that the earnings data is reported in 23 brackets. So the scatter plot shows 23 horizontal lines.

d.

```r
lm_height <- lm(earnings~height, eah)
summary(lm_height)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = eah)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47836 -21879  -7976  34323  50599
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -512.73    3386.86  -0.151     0.88
## height        707.67      50.49  14.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 26780 on 17868 degrees of freedom
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.01082
## F-statistic: 196.5 on 1 and 17868 DF,  p-value: < 2.2e-16
```

i.

The estimated slope is 707.67.

ii.

Note that the estimation model: $Y = -512.73 + 707.67X$. The estimated earning for 67 inches: -512.73 + 707.67*67 = 46,901.16. The estimated earning for 70 inches: -512.73 + 707.67*70 = 49,024.17. The estimated earning for 65 inches: -512.73 + 707.67*65 = 45,485.82.

e.

i.

Note that

$$X_{centimeter} = 2.54 * X_{inch} \rightarrow \frac{X_{centimeter}}{2.54} = X_{inch}$$

Thus the slope would be 707.67/2.54 = 278.61

ii.

The intercept would remain same as -512.73.

iii.

R-squared would remain same as 0.01088, because the error term is not affected.

iv.

Standard error of the regression would remain same, as 26780.

However, standard error of the estimated coefficient for height would be

$$SE(\beta_{centimeter}) = SE(\beta_{inch})/2.54 = 19.88$$

We can check!

```r
eah_in_centimeter <- eah %>%
  mutate(height_in_centimeter = height*2.54)

lm_height_in_centimeter <- lm(earnings~height_in_centimeter, eah_in_centimeter)
summary(lm_height_in_centimeter)
```

```
##
## Call:
## lm(formula = earnings ~ height_in_centimeter, data = eah_in_centimeter)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47836 -21879  -7976  34323  50599
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -512.73    3386.86  -0.151     0.88
## height_in_centimeter   278.61      19.88  14.016   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26780 on 17868 degrees of freedom
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.01082
## F-statistic: 196.5 on 1 and 17868 DF,  p-value: < 2.2e-16
```

f.

```r
eah_female <- eah %>% filter(sex == 0)

lm_height_female <- lm(earnings~height, eah_female)
summary(lm_height_female)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = eah_female)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42748 -22006  -7466  36641  46865
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12650.9     6383.7   1.982   0.0475 *
## height         511.2       98.9   5.169   2.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26800 on 9972 degrees of freedom
## Multiple R-squared:  0.002672,   Adjusted R-squared:  0.002572
## F-statistic: 26.72 on 1 and 9972 DF,  p-value: 2.396e-07
```

i.

The estimated slope is 511.2.

ii.

The predicted earning would be 511.2 more than the average.

g.

```r
eah_male <- eah %>% filter(sex == 1)

lm_height_male <- lm(earnings~height, eah_male)
summary(lm_height_male)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = eah_male)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50158 -22373  -8118  33091  59228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -43130.3     7068.5  -6.102  1.1e-09 ***
## height        1306.9      100.8  12.969  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26670 on 7894 degrees of freedom
## Multiple R-squared:  0.02086,    Adjusted R-squared:  0.02074
## F-statistic: 168.2 on 1 and 7894 DF,  p-value: < 2.2e-16
```

i.

The estimated slope is 1,306.9.

ii.

The predicted earning would be 1,306.9 more than the average.

h.

`Height` would be severly correlated with `sex`. `Height` also is expected to have correlation with `race`, `region`, `economic background` and so on. Thus the conditional mean of error term given Height would not be 0 in the situation.

---

3. EE5.1

a.

```r
summary(lm_height)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = eah)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -47836 -21879  -7976  34323  50599
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -512.73    3386.86  -0.151     0.88
## height        707.67      50.49  14.016   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26780 on 17868 degrees of freedom
## Multiple R-squared:  0.01088,    Adjusted R-squared:  0.01082
## F-statistic: 196.5 on 1 and 17868 DF,  p-value: < 2.2e-16
```

```r
sprintf(
  "CI for the slope is (%.4f, %.4f)",
  707.67 - 1.96*50.49,
  707.67 + 1.96*50.49
)
```

```
## [1] "CI for the slope is (608.7096, 806.6304)"
```

i.

The p-value for the slope is small enough (2e-16) to say significance.

ii.

CI for the slope is (608.7096, 806.6304).

b.

```r
summary(lm_height_female)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = eah_female)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42748 -22006  -7466  36641  46865
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12650.9     6383.7   1.982   0.0475 *
## height         511.2       98.9   5.169  2.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26800 on 9972 degrees of freedom
## Multiple R-squared:  0.002672,   Adjusted R-squared:  0.002572
## F-statistic: 26.72 on 1 and 9972 DF,  p-value: 2.396e-07
```

```r
sprintf(
  "CI for the slope is (%.4f, %.4f)",
  511.2 - 1.96*98.9,
  511.2 + 1.96*98.9
)
```

```
## [1] "CI for the slope is (317.3560, 705.0440)"
```

i.

The p-value for the slope is small enough (2.4e-07) to say significance.

ii.

CI for the slope is (317.3560, 705.0440).

c.

```r
summary(lm_height_male)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = eah_male)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -50158 -22373  -8118  33091  59228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -43130.3     7068.5  -6.102  1.1e-09 ***
## height        1306.9      100.8  12.969  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26670 on 7894 degrees of freedom
## Multiple R-squared:  0.02086,    Adjusted R-squared:  0.02074
## F-statistic: 168.2 on 1 and 7894 DF,  p-value: < 2.2e-16
```

```
sprintf(
  "CI for the slope is (%.4f, %.4f)",
  1306.9 - 1.96*100.8,
  1306.9 + 1.96*100.8
)
```

```
## [1] "CI for the slope is (1109.3320, 1504.4680)"
```

i.

The p-value for the slope is small enough (2e-16) to say significance.

ii.

CI for the slope is (1,109.3320, 1,504.4680).

d.

The null hypothesis is:

$$\beta_{female} = \beta_{male}$$

and the opposite hypothesis is:

$$\beta_{female} \neq \beta_{male}$$

```
diff <- 1306.9 - 511.2
se_diff <- sqrt(98.9^2 + 100.8^2)

(t_diff <- diff/se_diff)
```

```
## [1] 5.634646
```

The t-statistic is big enough (5.63) to say that the effect of height for men is different from that of women. It is also statistically significant to say that the effect of height for men is bigger than that of women.

e.

The occupations in which strength is unlikely to be important could be: 1 = Exec/Manager 2 = Professionals 3 = Technicians 5 = Administrat 12 = Precision production

Let's regress on the sample with the occupations above.

```
eah_intelli <- eah %>%
  filter(occupation %in% c(1, 2, 3, 5, 12))

lm_height_intelli <- lm(earnings~height, eah_intelli)
summary(lm_height_intelli)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = eah_intelli)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -52221 -23204  -7613  29328  41175
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4376.1     4733.9   0.924    0.355
## height         740.5       71.1  10.414   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26490 on 9500 degrees of freedom
## Multiple R-squared:  0.01129,    Adjusted R-squared:  0.01118
## F-statistic: 108.4 on 1 and 9500 DF,  p-value: < 2.2e-16
```

The effect of the height is still significant with p-value of 2e-16, and also its value is not that small. It seems the height of a worker is still important factor that determines his earning.

---

4. EE6.1

```r
head(bws)
```

```
## # A tibble: 6 x 12
##   nprevist alcohol tripre1 tripre2 tripre3 tripre0 birthweight smoker
##      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>       <dbl>  <dbl>
## 1       12       0       1       0       0       0        4253      1
## 2        5       0       0       1       0       0        3459      0
## 3       12       0       1       0       0       0        2920      1
## 4       13       0       1       0       0       0        2600      0
## 5        9       0       1       0       0       0        3742      0
## 6       11       0       1       0       0       0        3420      0
## # ... with 4 more variables: unmarried <dbl>, educ <dbl>, age <dbl>,
## #   drinks <dbl>
```

a.

```r
lm_smoker <- lm(birthweight~smoker, bws)
summary(lm_smoker)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker, data = bws)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3007.06  -313.06    26.94   366.94  2322.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3432.06      11.87 289.115   <2e-16 ***
```

```
## smoker       -253.23      26.95  -9.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 583.7 on 2998 degrees of freedom
## Multiple R-squared:  0.0286, Adjusted R-squared:  0.02828
## F-statistic: 88.28 on 1 and 2998 DF,  p-value: < 2.2e-16
```

The estimated effect is -253.23 on birthweight.

b.

```r
lm_bws_mul <- lm(birthweight~smoker+alcohol+nprevist, bws)
summary(lm_bws_mul)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker + alcohol + nprevist, data = bws)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2733.53  -307.57    21.42   358.09  2192.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3051.249     34.016  89.701  < 2e-16 ***
## smoker      -217.580     26.680  -8.155 5.07e-16 ***
## alcohol      -30.491     76.234  -0.400    0.689
## nprevist      34.070      2.855  11.933  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 570.5 on 2996 degrees of freedom
## Multiple R-squared:  0.07285,    Adjusted R-squared:  0.07192
## F-statistic: 78.47 on 3 and 2996 DF,  p-value: < 2.2e-16
```

i.

Alcohol and Nprevist are expected to be correlated with Smoker, and expected to determine the birthweight at the same time. This is the condition of the presence of the omitted variable bias.

ii.

The estimated effect in the excluded model is -253.23, and the effect in the included model is -217.580. That is, the estimated effect is about 15% different between the two models, which stands for the evidence of ommited variable bias.

iii.

$$\hat{Y} = 3051.249 - 217.580 * 1 - 30.491 * 0 + 34.070 * 8 = 3,106.229$$

iv.

Note that

$$\bar{R^2} = 1 - \frac{n-1}{n-k-1} * (1 - R^2)$$

As the sample size gets bigger, the two parameters gets closer. So the two parameters are similar with about 3,000 rows of sample.

v.

The interpretation of the coefficient is that the birthweight with one more prenatal visit is 34.070 bigger in average, holding other variables constant.

However, it's hard to say that prenatal visit has causal effect with the birthweight. Actually the number of prenatal visits works as control variable to dismiss correlation that causes ommited variable bias.

c.

```r
x_on_others <- lm(smoker~alcohol+nprevist, bws)
y_on_others <- lm(birthweight~alcohol+nprevist, bws)

resid <- tibble(
  x_resid = residuals(x_on_others),
  y_resid = residuals(y_on_others)
)

lm_frisch_waugh <- lm(y_resid~x_resid, resid)
summary(lm_frisch_waugh)
```

```
##
## Call:
## lm(formula = y_resid ~ x_resid, data = resid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2733.53  -307.57    21.42   358.09  2192.70
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.346e-13  1.041e+01   0.000        1
## x_resid     -2.176e+02  2.667e+01  -8.158 4.95e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 570.3 on 2998 degrees of freedom
## Multiple R-squared:  0.02172,    Adjusted R-squared:  0.02139
## F-statistic: 66.55 on 1 and 2998 DF,  p-value: 4.955e-16
```

We can see that -2.176e+02 ~= -217.580.

d.

```r
lm_tripre <- lm(birthweight~smoker+alcohol+tripre0+tripre2+tripre3, bws)
summary(lm_tripre)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker + alcohol + tripre0 + tripre2 +
##     tripre3, data = bws)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -3029.55  -307.55    31.35   372.45  2401.29 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  3454.55      12.65 273.077  < 2e-16 ***
## smoker       -228.85      27.16  -8.424  < 2e-16 ***
## alcohol       -15.10      77.54  -0.195 0.845613    
## tripre0      -697.97     106.88  -6.531 7.66e-11 ***
## tripre2      -100.84      29.62  -3.404 0.000672 ***
## tripre3      -136.96      59.58  -2.299 0.021595 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 578.7 on 2994 degrees of freedom
## Multiple R-squared:  0.04647,    Adjusted R-squared:  0.04487 
## F-statistic: 29.18 on 5 and 2994 DF,  p-value: < 2.2e-16
```

```r
lm_tripre_temp <- lm(birthweight~smoker+alcohol+tripre0+tripre1+tripre2+tripre3, bws)
summary(lm_tripre_temp)
```

```
## 
## Call:
## lm(formula = birthweight ~ smoker + alcohol + tripre0 + tripre1 + 
##     tripre2 + tripre3, data = bws)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -3029.55  -307.55    31.35   372.45  2401.29 
## 
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  3317.59      59.00  56.231  < 2e-16 ***
## smoker       -228.85      27.16  -8.424  < 2e-16 ***
## alcohol       -15.10      77.54  -0.195   0.8456    
## tripre0      -561.01     120.88  -4.641 3.61e-06 ***
## tripre1       136.96      59.58   2.299   0.0216 *  
## tripre2        36.12      64.17   0.563   0.5736    
## tripre3           NA         NA      NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 578.7 on 2994 degrees of freedom
## Multiple R-squared:  0.04647,    Adjusted R-squared:  0.04487 
## F-statistic: 29.18 on 5 and 2994 DF,  p-value: < 2.2e-16
```

i.

Coeffifiecnt of Tripre3 cannot be calculated because of perfect multicollinearity. Note that Note that Tripre0 + Tripre1 + Tripre2 + Tripre3 = 1. To avoid the multicollinearity, one of the four variables should be excluded.

ii.

Tripre0 == 1 means that the mother has never had prenatal visit, which decreases the birthweight by 697.97 in average holding other variables constant. Specifically, the value means the mean difference of birthweight of the mother with 0 prenatal visit from that of 1 prenatal visit.

iii.

Each value stands for the mean difference of birthweight of the mother with 2/3 prenatal visits from that of 1.

iv.

Let's compare the squared R of the regressions. The squared R for regression in (b) is 0.07285, and for regression in (d) is 0.04647. The regression in (b) explains a larger fraction of the variance tha the regression in (d).

---

5. EE7.1

Regression (1)

```r
summary(lm_smoker)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker, data = bws)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3007.06  -313.06    26.94   366.94  2322.94
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3432.06      11.87 289.115   <2e-16 ***
## smoker       -253.23      26.95  -9.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 583.7 on 2998 degrees of freedom
## Multiple R-squared:  0.0286, Adjusted R-squared:  0.02828
## F-statistic: 88.28 on 1 and 2998 DF,  p-value: < 2.2e-16
```

Regression (2)

```r
summary(lm_bws_mul)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker + alcohol + nprevist, data = bws)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2733.53  -307.57    21.42   358.09  2192.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3051.249     34.016  89.701  < 2e-16 ***
## smoker      -217.580     26.680  -8.155 5.07e-16 ***
## alcohol      -30.491     76.234  -0.400    0.689
## nprevist      34.070      2.855  11.933  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 570.5 on 2996 degrees of freedom
## Multiple R-squared:  0.07285,    Adjusted R-squared:  0.07192
## F-statistic: 78.47 on 3 and 2996 DF,  p-value: < 2.2e-16
```

Regression (3)

```r
lm_bws_mul2 <- lm(birthweight~smoker+alcohol+nprevist+unmarried, bws)
summary(lm_bws_mul2)
```

```
##
## Call:
## lm(formula = birthweight ~ smoker + alcohol + nprevist + unmarried,
##     data = bws)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2798.81  -309.22    25.37   361.80  2363.70
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3134.400     35.656  87.907  < 2e-16 ***
## smoker      -175.377     27.099  -6.472 1.13e-10 ***
## alcohol      -21.083     75.607  -0.279     0.78
## nprevist      29.603      2.898  10.213  < 2e-16 ***
## unmarried   -187.133     26.007  -7.195 7.84e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 565.7 on 2995 degrees of freedom
## Multiple R-squared:  0.08861,    Adjusted R-squared:  0.08739
## F-statistic: 72.79 on 4 and 2995 DF,  p-value: < 2.2e-16
```

a.

> Regression (1): -253.23 / se=26.95 Regression (2): -217.580 / se=26.680 Regression (3): -175.377 / se=27.099

b.

```r
sprintf(
  "CI for regression (1) is (%.4f, %.4f)",
  -253.23 - 1.96*26.95,
  -253.23 + 1.96*26.95
)
```

```
## [1] "CI for regression (1) is (-306.0520, -200.4080)"
```

```r
sprintf(
  "CI for regression (2) is (%.4f, %.4f)",
  -217.580 - 1.96*26.680,
  -217.580 + 1.96*26.680
)
```

```
## [1] "CI for regression (2) is (-269.8728, -165.2872)"
```

```r
sprintf(
  "CI for regression (3) is (%.4f, %.4f)",
  -175.377 - 1.96*27.099,
  -175.377 + 1.96*27.099
)
```

```
## [1] "CI for regression (3) is (-228.4910, -122.2630)"
```

c.

Smoker would be correlated with Alcohol, which is expected to determine the birthweight. Thus, the regression coefficient would suffer from omitted variable bias.

d.

Unmarried could be correlated with Smoker, and would determine the birthweight in indirect ways. For example, unmarried parent would have bad economic situation, lack of care, and so on. So the regression coefficient would suffer from omitted variable bias.

e.

i.

```r
sprintf(
  "CI is (%.4f, %.4f)",
  -187.133 - 1.96*26.007,
  -187.133 + 1.96*26.007
)
```

```
## [1] "CI is (-238.1067, -136.1593)"
```

ii.

As the confidence interval does not contain 0, the coefficient is statistically significant.

iii.

The magnitude of the coefficient is even bigger than the coefficient of Smoker. Noting that the sample standard deviation of the sample is 592.1629, Unmarried affect 0.3 standard deviation of the birthweight. It is large.

```r
sd(bws[["birthweight"]])
```

```
## [1] 592.1629
```

iv.

Unmarried in the regression is used as control variable which has a correlation with the error term, to dismiss the omitted variable bias. Thus the coefficient would be biased, and cannot be interpreted as causal effect.

f.

```r
regressor_factor <- c(
  "smoker",
  "nprevist",
  "unmarried",
  "alcohol",
  "educ",
  "age",
```

```
  "(Intercept)",
  "-",
  "R-adj",
  "n"
)

make_reg_table <- function (reg) {

  summ <- summary(reg)
  summ_coef <- summ$coefficients

  n_reg <- length(summ_coef[, 1:2])/2

  coeffs <- c()

  for (i in 1:n_reg) {
    coef <- summ_coef[i]
    se_coef <- summ_coef[n_reg + i]

    t_val <- abs(coef/se_coef)

    if (t_val >= 2.58) {
      sf <- "***"
    } else if (t_val >= 1.96) {
      sf <- "**"
    } else if (t_val >= 1.645) {
      sf <- "*"
    } else {
      sf <- ""
    }

    coeffs <- c(coeffs, sprintf("%.1f (%.1f) %s", coef, se_coef, sf))
  }

  tibble(
    regressor = factor(c(dimnames(summ_coef)[[1]], "-", "R-adj", "n"), regressor_factor),
    value = c(coeffs, "-", round(summ$adj.r.squared, 4), 3000)
  )
}

reg1 <- lm(birthweight~smoker, bws)
reg2 <- lm(birthweight~smoker+nprevist+unmarried, bws)
reg3 <- lm(birthweight~smoker+nprevist+unmarried+alcohol, bws)
reg4 <- lm(birthweight~smoker+nprevist+unmarried+educ, bws)
reg5 <- lm(birthweight~smoker+nprevist+unmarried+age, bws)
reg6 <- lm(birthweight~smoker+nprevist+unmarried+alcohol+educ+age, bws)

for (i in 1:6) {
  assign(
    sprintf("reg%i_table", i),
    make_reg_table(get(sprintf("reg%i", i)))
  )
}
```

```r
reg1_table %>%
  full_join(
    reg2_table,
    by = c("regressor")
  ) %>%
  full_join(
    reg3_table,
    by = c("regressor")
  ) %>%
  full_join(
    reg4_table,
    by = c("regressor")
  ) %>%
  full_join(
    reg5_table,
    by = c("regressor")
  ) %>%
  full_join(
    reg6_table,
    by = c("regressor")
  ) %>%
  arrange(
    regressor
  ) %>%
  rename(
    "(1)" = value.x,
    "(2)" = value.y,
    "(3)" = value.x.x,
    "(4)" = value.y.y,
    "(5)" = value.x.x.x,
    "(6)" = value.y.y.y
  )
```

```
## # A tibble: 10 x 7
##    regressor  `(1)`      `(2)`      `(3)`     `(4)`     `(5)`     `(6)`
##    <fct>      <chr>      <chr>      <chr>     <chr>     <chr>     <chr>
##  1 smoker     -253.2 (2~ -176.2 (2~ -175.4 (~ -177.8 (~ -177.7 (~ -177.0 (~
##  2 nprevist   <NA>       29.6 (2.9~ 29.6 (2.~ 29.8 (2.~ 29.8 (2.~ 29.8 (2.~
##  3 unmarried  <NA>       -187.3 (2~ -187.1 (~ -190.0 (~ -199.7 (~ -199.3 (~
##  4 alcohol    <NA>       <NA>       "-21.1 (~ <NA>      <NA>      "-14.8 (~
##  5 educ       <NA>       <NA>       <NA>      "-1.9 (5~ <NA>      "0.2 (5.~
##  6 age        <NA>       <NA>       <NA>      <NA>      "-2.5 (2~ "-2.5 (2~
##  7 (Intercep~ 3432.1 (1~ 3134.0 (3~ 3134.4 (~ 3158.3 (~ 3202.0 (~ 3199.4 (~
##  8 -          -          -          -         -         -         -
##  9 R-adj      0.0283     0.0877     0.0874    0.0874    0.0878    0.0872
## 10 n          3000       3000       3000      3000      3000      3000
```

Other variables like `alcohol`, `educ`, `age` seem to have insignificant affect on birthweight. And taking a look at the first row, the coefficient of `smoker` is about -177~-176 for regression (2)~(6), which is robust. Thus, the confidence interval for regression (3) in (b), which is (-228.4910, -122.2630) seems to be reasonable.