Logistic Regression

Matthew Seguin

3.

```
library(rio)
cps_98 <- import("cps98_short.xls")
```

a.

```
reg1 <- lm(ahe ~ age + female + bachelor, data = cps_98)
summary(reg1)
```

```
##
## Call:
## lm(formula = ahe ~ age + female + bachelor, data = cps_98)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.105  -4.625  -1.087   3.377  31.724
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.26513    2.43807   1.749 0.080540 .
## age          0.29119    0.08029   3.627 0.000302 ***
## female      -2.11128    0.45523  -4.638    4e-06 ***
## bachelor     5.50795    0.44486  12.381  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.875 on 970 degrees of freedom
## Multiple R-squared:  0.1513, Adjusted R-squared:  0.1487
## F-statistic: 57.64 on 3 and 970 DF,  p-value: < 2.2e-16
```

We just ran the regression $ahe = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 bachelor + U$.
Under this model if all other variables are held constant and age increases by 1
(i.e. $age_1 = age_0 + 1$) we will get the following:

$$\Delta ahe = ahe_1 - ahe_0 = (\beta_0 + \beta_1 age_1 + \beta_2 female + \beta_3 bachelor) - (\beta_0 + \beta_1 age_0 + \beta_2 female + \beta_3 bachelor) = \beta_1 (age_1 - age_0) = \beta_1$$

Therefore the expected change in ahe for age going from 25 to 26 is the same as that of age going from 33 to 34 since both are just increases by 1.
As per above we have that in both cases $\Delta ahe = \beta_1 \approx \hat{\beta}_1$.
This was estimated via the results of our regression using OLS, the result ($\hat{\beta}_1$) is shown below.

```
as.numeric(reg1$coefficients[2])
```

```
## [1] 0.2911944
```

So the estimated change in average hourly earnings as a result of age going from 25 to 26 or age going from 33 to 34, holding all other variables constant, is an increase of 0.2911944 which is \$0.2911944 in the real world.

b.

```
reg2 <- lm(log(ahe) ~ age + female + bachelor, data = cps_98)
summary(reg2)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + female + bachelor, data = cps_98)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9466 -0.3062  0.0275  0.3190  1.4150
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.879749   0.164949  11.396  < 2e-16 ***
## age          0.018909   0.005432   3.481 0.000522 ***
## female      -0.139088   0.030799  -4.516 7.08e-06 ***
## bachelor     0.376630   0.030097  12.514  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4651 on 970 degrees of freedom
## Multiple R-squared:  0.1524, Adjusted R-squared:  0.1498
## F-statistic: 58.13 on 3 and 970 DF,  p-value: < 2.2e-16
```

We just ran the regression $log(ahe) = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 bachelor + U$.
Under this model if all other variables are held constant and age increases by 1 (i.e. $\Delta age = 1$) we will get the following:

$$log(ahe+\Delta ahe) - log(ahe) = (\beta_0+\beta_1(age+\Delta age)+\beta_2 female+\beta_3 bachelor) - (\beta_0+\beta_1 age+\beta_2 female+\beta_3 bachelor) = \beta_1(\Delta age) = \beta_1$$

Now recall that $log(x + \Delta x) - log(x) \approx \frac{\Delta x}{x}$.
So in our case we have that:

$$\frac{\Delta ahe}{ahe} \approx log(ahe + \Delta ahe) - log(ahe) = \beta_1$$

Therefore an increase in age by 1 causes an expected increase in ahe by about $100\beta_1\% \approx 100\hat{\beta}_1\%$.
This can be estimated via the results of our regression using OLS, the results are shown below.

```
as.numeric(reg2$coefficients[2])
```

```
## [1] 0.01890923
```

```
100*as.numeric(reg2$coefficients[2])
```

```
## [1] 1.890923
```

So the estimated change in average hourly earnings as a result of age going from 25 to 26 or age going from 33 to 34, holding all other variables constant, is an increase in average hourly earnings by about 1.890923%.

If we want to find the estimated total change in *ahe* we may use the following process which I have on the next page.

When age increases by 1 (i.e. $\Delta age = 1$) the expected change in ahe is given by:

$$\Delta ahe \approx \beta_1 ahe \approx \hat{\beta}_1 \hat{ahe}$$

This can be estimated via the results of our regression using OLS. If we use the desired value for age and mean of the other regressors then we can approximate $\hat{ahe}$ for a given age:

$$log(\hat{ahe}) \approx log(\hat{ahe}) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 \mathbb{E}(female) + \hat{\beta}_3 \mathbb{E}(bachelor)$$

From which it follows:

$$\hat{ahe} = e^{log(\hat{ahe})} \approx e^{log(\hat{ahe})} = e^{\hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 \mathbb{E}(female) + \hat{\beta}_3 \mathbb{E}(bachelor)}$$

The means of the regressors from our data are shown below.

```
b_0_hat <- as.numeric(reg2$coefficients[1])
b_1_hat <- as.numeric(reg2$coefficients[2])
b_2_hat <- as.numeric(reg2$coefficients[3])
b_3_hat <- as.numeric(reg2$coefficients[4])
E_bachelor <- mean(cps_98$bachelor)
E_bachelor
```

```
## [1] 0.4784394
```

```
E_female <- mean(cps_98$female)
E_female
```

```
## [1] 0.389117
```

- For age going from 25 to 26:

Under our model:

$$log(\hat{ahe}) \approx \hat{\beta}_0 + 25\hat{\beta}_1 + \hat{\beta}_2 \mathbb{E}(female) + \hat{\beta}_3 \mathbb{E}(bachelor)$$

The result of which is shown below.

```
log_ahe_hat_25 <- b_0_hat + 25*b_1_hat + b_2_hat*E_female + b_3_hat*E_bachelor
ahe_hat_25 <- exp(log_ahe_hat_25)
delta_ahe_hat_25 <- b_1_hat*ahe_hat_25
delta_ahe_hat_25
```

```
## [1] 0.2254735
```

So the estimated change in average hourly earnings as a result of age going from 25 to 26, holding all other variables constant, is an increase of 0.2254735 which is \$0.2254735 in the real world.

- For age going from 33 to 34:

Under our model:

$$log(\hat{ahe}) \approx \hat{\beta}_0 + 33\hat{\beta}_1 + \hat{\beta}_2 \mathbb{E}(female) + \hat{\beta}_3 \mathbb{E}(bachelor)$$

The result of which is shown below.

```r
log_ahe_hat_33 <- b_0_hat + 33*b_1_hat + b_2_hat*E_female + b_3_hat*E_bachelor
ahe_hat_33 <- exp(log_ahe_hat_33)
delta_ahe_hat_33 <- b_1_hat*ahe_hat_33
delta_ahe_hat_33
```

```
## [1] 0.2622967
```

So the estimated change in average hourly earnings as a result of age going from 33 to 34, holding all other variables constant, is an increase of 0.2622967 which is \$0.2622967 in the real world.

c.

```
reg3 <- lm(log(ahe) ~ log(age) + female + bachelor, data = cps_98)
summary(reg3)
```

```
##
## Call:
## lm(formula = log(ahe) ~ log(age) + female + bachelor, data = cps_98)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94873 -0.30574  0.02728  0.31969  1.41860
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5723     0.5445   1.051 0.293464
## log(age)      0.5519     0.1600   3.448 0.000588 ***
## female       -0.1391     0.0308  -4.515 7.12e-06 ***
## bachelor      0.3763     0.0301  12.504  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4652 on 970 degrees of freedom
## Multiple R-squared:  0.1522, Adjusted R-squared:  0.1496
## F-statistic: 58.04 on 3 and 970 DF,  p-value: < 2.2e-16
```

We just ran the regression $log(ahe) = \beta_0 + \beta_1 log(age) + \beta_2 female + \beta_3 bachelor + U$.
Under this model if all other variables are held constant and age increases by 1 (i.e. $\Delta age = 1$) we will get the following:

$$log(ahe+\Delta ahe) - log(ahe) = (\beta_0 + \beta_1 log(age+\Delta age) + \beta_2 female + \beta_3 bachelor) - (\beta_0 + \beta_1 log(age) + \beta_2 female + \beta_3 bachelor)$$

$$= \beta_1 \big(log(age + \Delta age) - log(age)\big)$$

Now recall that $log(x + \Delta x) - log(x) \approx \frac{\Delta x}{x}$.
So in our case we have that:

$$\frac{\Delta ahe}{ahe} \approx log(ahe + \Delta ahe) - log(ahe) = \beta_1 \big(log(age + \Delta age) - log(age)\big) \approx \beta_1 \frac{\Delta age}{age} = \beta_1 \frac{1}{age}$$

Therefore an increase in age by 1 causes an expected increase in ahe by about $100\frac{\beta_1}{age}\% \approx 100\frac{\hat{\beta}_1}{age}\%$.
This can be estimated via the results of our regression using OLS, the results are shown below.

- For age going from 25 to 26:

```
100*as.numeric(reg3$coefficients[2])/25
```

```
## [1] 2.207615
```

So the estimated change in average hourly earnings as a result of age going from 25 to 26, holding all other variables constant, is an increase in average hourly earnings by about 2.207615%.

- For age going from 33 to 34:

```
100*as.numeric(reg3$coefficients[2])/33
```

```
## [1] 1.672436
```

So the estimated change in average hourly earnings as a result of age going from 33 to 34, holding all other variables constant, is an increase in average hourly earnings by about 1.672436%.

If we want to find the estimated total change in *ahe* we may use the following process which I have on the next page.

When age increases by 1 (i.e. $\Delta age = 1$) the expected change in ahe is given by:

$$\Delta ahe \approx \beta_1 \frac{ahe}{age} \approx \hat{\beta}_1 \frac{\hat{ahe}}{age}$$

This can be estimated via the results of our regression using OLS. If we use the desired value for age and mean of the other regressors then we can approximate $\hat{ahe}$ for a given age:

$$log(\hat{ahe}) \approx log(\hat{ahe}) = \hat{\beta}_0 + \hat{\beta}_1 log(age) + \hat{\beta}_2 \mathbb{E}(female) + \hat{\beta}_3 \mathbb{E}(bachelor)$$

From which it follows:

$$\hat{ahe} = e^{log(\hat{ahe})} \approx e^{log(\hat{ahe})} = e^{\hat{\beta}_0 + \hat{\beta}_1 log(age) + \hat{\beta}_2 \mathbb{E}(female) + \hat{\beta}_3 \mathbb{E}(bachelor)}$$

The means of the other regressors were found before here they are displayed again.

```
b_0_hat <- as.numeric(reg3$coefficients[1])
b_1_hat <- as.numeric(reg3$coefficients[2])
b_2_hat <- as.numeric(reg3$coefficients[3])
b_3_hat <- as.numeric(reg3$coefficients[4])
E_bachelor
```

```
## [1] 0.4784394
```

```
E_female
```

```
## [1] 0.389117
```

- For age going from 25 to 26:

Under our model:

$$log(\hat{ahe}) \approx \hat{\beta}_0 + \hat{\beta}_1 log(25) + \hat{\beta}_2 \mathbb{E}(female) + \hat{\beta}_3 \mathbb{E}(bachelor)$$

The result of which is shown below.

```
log_ahe_hat_25 <- b_0_hat + log(25)*b_1_hat + b_2_hat*E_female + b_3_hat*E_bachelor
ahe_hat_25 <- exp(log_ahe_hat_25)
delta_ahe_hat_25 <- b_1_hat*ahe_hat_25/25
delta_ahe_hat_25
```

```
## [1] 0.2622447
```

So the estimated change in average hourly earnings as a result of age going from 25 to 26, holding all other variables constant, is an increase of 0.2622447 which is \$0.2622447 in the real world.

- For age going from 33 to 34:

Under our model:

$$log(\hat{ahe}) \approx \hat{\beta}_0 + \hat{\beta}_1 log(33) + \hat{\beta}_2 \mathbb{E}(female) + \hat{\beta}_3 \mathbb{E}(bachelor)$$

The result of which is shown below.

```
log_ahe_hat_33 <- b_0_hat + log(33)*b_1_hat + b_2_hat*E_female + b_3_hat*E_bachelor
ahe_hat_33 <- exp(log_ahe_hat_33)
delta_ahe_hat_33 <- b_1_hat*ahe_hat_33/33
delta_ahe_hat_33
```

```
## [1] 0.2315677
```

So the estimated change in average hourly earnings as a result of age going from 33 to 34, holding all other variables constant, is an increase of 0.2315677 which is \$0.2315677 in the real world.

d.

To answer the question of whether the effect of age on average hourly earnings is different for men and women we need to add another variable to our regression that includes information about both simultaneously. It is fairly intuitive to make the variable $agegen = (age)(female)$, the product of the variables age and female, then run a previous regression adding the new variable. I will run the regression of $log(ahe)$ using age, $female$, $bachelor$, and $agegen$ as regressors, I don't want to use $log(age)$ because I want to see the effect of just age in comparison to gender.

```
agegen <- cps_98$age*cps_98$female
cps_98$agegen <- agegen
reg4 <- lm(log(ahe) ~ age + female + bachelor + agegen, data = cps_98)
summary(reg4)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + female + bachelor + agegen, data = cps_98)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94601 -0.30038  0.02286  0.32080  1.44738
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.678257   0.209828   7.998 3.58e-15 ***
## age          0.025672   0.006961   3.688 0.000239 ***
## female       0.373841   0.331963   1.126 0.260379
## bachelor     0.374657   0.030102  12.446  < 2e-16 ***
## agegen      -0.017222   0.011098  -1.552 0.121031
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4648 on 969 degrees of freedom
## Multiple R-squared:  0.1545, Adjusted R-squared:  0.151
## F-statistic: 44.26 on 4 and 969 DF,  p-value: < 2.2e-16
```

We just ran the regression $log(ahe) = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 bachelor + \beta_4 age \times female + U$.
Under this model if a person is female ($female = 1$) then we get:

$$log(ahe) = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 bachelor + \beta_4 age \times female = \beta_0 + \beta_1 age + \beta_2 + \beta_3 bachelor + \beta_4 age$$

$$= (\beta_0 + \beta_2) + (\beta_1 + \beta_4) age + \beta_3 bachelor$$

So the effect on $log(ahe)$ as a result of age is given by $\beta_1 + \beta_4$ when a person is female (i.e. $female = 1$).
Under this model if a person is female ($female = 0$) then we get:

$$log(ahe) = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 bachelor + \beta_4 age \times female = \beta_0 + \beta_1 age + \beta_3 bachelor$$

So the effect on $log(ahe)$ as a result of age is given by β_1 when a person is male (i.e. $female = 0$).

Therefore if we want to test if the effect of age on average hourly earnings is different for men and women we can test the simple hypothesis $H_0 : \beta_4 = 0$.

I will perform this test on the next page.

We know the OLS estimate from our regression and we know the coefficient is approximately normal so we can find the p-value with the t-statistic.

$$t_{H_0} = \frac{\hat{\beta}_4 - 0}{SE(\hat{\beta}_4)} = \frac{\hat{\beta}_4 - 0}{SE(\hat{\beta}_4)} \sim N(0,1)$$

We can compute this manually or just get the result from our regression, I will show both below.

```
reg4_data <- summary(reg4)
b_4_hat <- reg4_data$coefficients[5, "Estimate"]
b_4_hat
```

```
## [1] -0.01722202
```

```
se_b_4_hat <- reg4_data$coefficients[5, "Std. Error"]
se_b_4_hat
```

```
## [1] 0.01109792
```

```
t_calc <- b_4_hat / se_b_4_hat
t_calc
```

```
## [1] -1.551823
```

```
t_from_reg <- reg4_data$coefficients[5, "t value"]
t_from_reg
```

```
## [1] -1.551823
```

As you can see both t statistics are the same, the calculated one and the one just extracted from the regression. I will now calculate the p-value and give a conclusion. We can compute the p-value directly or just extract it from the regression, again I will show both below.

```
p_calc <- 2*pnorm(-abs(t_calc))
p_calc
```

```
## [1] 0.1207045
```

```
p_from_reg <- reg4_data$coefficients[5, "Pr(>|t|)"]
p_from_reg
```

```
## [1] 0.121031
```

As you can see these are both very close, the reason they are not exactly the same is because I used a normal approximation while R used the exact p-value from the t distribution.

In either case we see the p-value $> \alpha = 0.05$ so we do not reject $H_0 : \beta_4 = 0$.
That is we can say that the effect of age on average hourly earnings is NOT different for men and women.

e.

Similar to the previous part, to answer the question of whether the effect of age on average hourly earnings is different for high school graduates and college graduates we need to add another variable to our regression that includes information about both simultaneously. It is fairly intuitive to make the variable $agecol = (age)(bachelor)$, the product of the variables age and bachelor, then run a previous regression adding the new variable. I will run the regression of $log(ahe)$ using age, $female$, $bachelor$, and $agecol$ as regressors, I don't want to use $log(age)$ because I want to see the effect of just age in comparison to education.

```
agecol <- cps_98$age*cps_98$bachelor
cps_98$agecol <- agecol
reg5 <- lm(log(ahe) ~ age + female + bachelor + agecol, data = cps_98)
summary(reg5)
```

```
##
## Call:
## lm(formula = log(ahe) ~ age + female + bachelor + agecol, data = cps_98)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.94668 -0.30695  0.02743  0.31946  1.41349
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.8887047  0.2227994   8.477  < 2e-16 ***
## age          0.0186097  0.0073887   2.519   0.0119 *
## female      -0.1390027  0.0308475  -4.506 7.41e-06 ***
## bachelor     0.3571663  0.3266751   1.093   0.2745
## agecol       0.0006528  0.0109100   0.060   0.9523
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4654 on 969 degrees of freedom
## Multiple R-squared:  0.1524, Adjusted R-squared:  0.1489
## F-statistic: 43.55 on 4 and 969 DF,  p-value: < 2.2e-16
```

We just ran the regression $log(ahe) = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 bachelor + \beta_4 age \times bachelor + U$.
Under this model if a person is a college graduate ($bachelor = 1$) then we get:

$$log(ahe) = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 bachelor + \beta_4 age \times bachelor = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 + \beta_4 age$$

$$= (\beta_0 + \beta_3) + (\beta_1 + \beta_4) age + \beta_2 bachelor$$

So the effect on $log(ahe)$ as a result of age is given by $\beta_1 + \beta_4$ when a person is a college graduate (i.e. $bachelor = 1$).
Under this model if a person is only a high school graduate ($bachelor = 0$) then we get:

$$log(ahe) = \beta_0 + \beta_1 age + \beta_2 female + \beta_3 bachelor + \beta_4 age \times bachelor = \beta_0 + \beta_1 age + \beta_2 female$$

So the effect on $log(ahe)$ as a result of age is given by β_1 when a person is only a high school graduate (i.e. $bachelor = 0$).

Therefore if we want to test if the effect of age on average hourly earnings is different for high school graduates and college graduates we can test the simple hypothesis $H_0 : \beta_4 = 0$.

I will perform this test on the next page (it is a similar process to the previous part).

We know the OLS estimate from our regression and we know the coefficient is approximately normal so we can find the p-value with the t-statistic.

$$t_{H_0} = \frac{\hat{\beta}_4 - 0}{SE(\hat{\beta}_4)} = \frac{\hat{\beta}_4 - 0}{SE(\hat{\beta}_4)} \sim N(0,1)$$

We can compute this manually or just get the result from our regression, I will show both below.

```
reg5_data <- summary(reg5)
b_4_hat <- reg5_data$coefficients[5, "Estimate"]
b_4_hat
```

```
## [1] 0.0006527996
```

```
se_b_4_hat <- reg5_data$coefficients[5, "Std. Error"]
se_b_4_hat
```

```
## [1] 0.01090996
```

```
t_calc <- b_4_hat / se_b_4_hat
t_calc
```

```
## [1] 0.05983518
```

```
t_from_reg <- reg5_data$coefficients[5, "t value"]
t_from_reg
```

```
## [1] 0.05983518
```

As you can see both t statistics are the same, the calculated one and the one just extracted from the regression. I will now calculate the p-value and give a conclusion. We can compute the p-value directly or just extract it from the regression, again I will show both below.

```
p_calc <- 2*pnorm(-abs(t_calc))
p_calc
```

```
## [1] 0.9522869
```

```
p_from_reg <- reg5_data$coefficients[5, "Pr(>|t|)"]
p_from_reg
```

```
## [1] 0.9522992
```

As you can see these are both very close, the reason they are not exactly the same is because I used a normal approximation while R used the exact p-value from the t distribution.

In either case we see the p-value $> \alpha = 0.05$ so we do not reject $H_0 : \beta_4 = 0$.
That is we can say that the effect of age on average hourly earnings is NOT different for high school graduates and college graduates.