SMML Class 3 Lab

John Kubale

9/10/2024

Import Income2.csv using either the read.csv() or read_csv() function (available on Canvas as well as JWHT Website) and explore the data.

```
dat <- read_csv("~/UMD/classes/stat_mod_ML_1_SURV615/class_3/Income2.csv")
glimpse(dat)
```

```
## Rows: 30
## Columns: 4
## $ ...1      <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1~
## $ Education <dbl> 21.58621, 18.27586, 12.06897, 17.03448, 19.93103, 18.27586, ~
## $ Seniority <dbl> 113.10345, 119.31034, 100.68966, 187.58621, 20.00000, 26.206~
## $ Income    <dbl> 99.91717, 92.57913, 34.67873, 78.70281, 68.00992, 71.50449, ~
```

Treat `Income` as Y and work on the following questions.

1. What is the mean of `Income`?

- The mean of income is 62.7.

```
dat |> summarise(mean(Income), sd(Income))
```

```
## # A tibble: 1 x 2
##   `mean(Income)` `sd(Income)`
##            <dbl>        <dbl>
## 1           62.7         27.0
```

2. Examine `Income` in linear regression with no predictor and save the model as an object called "no_pred". Use the summary function to examine the model.

```
no_pred <- lm(Income~1, data=dat)
summary(no_pred)
```

```
##
```

```
## Call:
## lm(formula = Income ~ 1, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45.13 -26.35   8.06  23.19  37.17
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   62.745      4.932   12.72 2.16e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.01 on 29 degrees of freedom
```

A. What parameter are you estimating in the model? How would you interpret the estimate? How would you put these into a formula?

- We are estimating the Intercept denoted as β_0, and we estimate the expected average value of Income as 62.744733 and we can write it as $\overline{Income_i} = \beta_0 + \epsilon_i$.

B. What are the residuals and what do they represent? How can you extract the estimates below "Coefficients" when you run summary(no_pred)?

- The residuals are the differences between the observed and expected value, unexplained error not explained by our model. Residuals are centered at zero and have a range of $-.45 and .37$ in this model.

C. How does results from the regression model compare to the mean in #1?

- When there are no explanatory variables in the model, our intercept is our expected value and in this case the observed mean is our best "guess" for predictions.

3. Examine `Income` as a fuction of `Education` in linear regression. Explore the results using the summary() function and by extracting the regression coefficients (together and separately).

```
mod_edu <- lm(Income ~ Education, data = dat)
summary(mod_edu)
```

```
##
## Call:
## lm(formula = Income ~ Education, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.568  -8.012   1.474   5.754  23.701
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -41.9166     9.7689  -4.291 0.000192 ***
## Education     6.3872     0.5812  10.990 1.15e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.93 on 28 degrees of freedom
## Multiple R-squared:  0.8118, Adjusted R-squared:  0.8051
## F-statistic: 120.8 on 1 and 28 DF,  p-value: 1.151e-11
```

A. What is this type of linear regression model called? How would you put this into a formula?

- This is a simple linear regression model with one explanatory variable and we would express this as $\overline{Income_I} = \beta_0 + \beta_1 \times Education_i + \epsilon_i$

B. Is the result the same as the one from #2?

- They are not the same as we now specify that income is a function of education and thus have a slope in our model.

4. Examine `Income` as a fuction of `Education` with no intercept in linear regression.

```
mod_edu_no_int <- lm(Income ~ Education + 0, data = dat)
summary(mod_edu_no_int)
```

```
##
## Call:
## lm(formula = Income ~ Education + 0, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.134 -12.512   1.409   8.968  22.343
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## Education   3.9562     0.1639   24.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.09 on 29 degrees of freedom
## Multiple R-squared:  0.9526, Adjusted R-squared:  0.951
## F-statistic: 582.8 on 1 and 29 DF,  p-value: < 2.2e-16
```

A. How would you put this into a formula?

- We would express this model as $\overline{Income_i} = \beta_1 \times Education_i + \epsilon_i$ as we drop the intercept parameter.

B. Is the result the same as the one from #3? Do you have concerns about this modeling approach?

- The results are not the same as we dropped the intercept and kept the parameter for education. I am concerned about dropping the intercept in this case as the interpretability of this model is reduced and offers us no value.

5. Examine `Income` as a function of `Seniority` in linear regression

```
mod_sen <- lm(Income ~ Seniority, data = dat)
summary(mod_sen)
```

```
##
## Call:
## lm(formula = Income ~ Seniority, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.764 -20.232   7.925  20.686  34.622
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.15833    8.51594   4.598 8.31e-05 ***
## Seniority    0.25129    0.07836   3.207  0.00335 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.51 on 28 degrees of freedom
## Multiple R-squared:  0.2686, Adjusted R-squared:  0.2425
## F-statistic: 10.28 on 1 and 28 DF,  p-value: 0.003347
```

A. How would you put this into a formula?

- We would express this model as $\overline{Income_i} = \beta_0 + \beta_1 \times Seniority_i + \epsilon_i$

B. How does this compare to the result in #3? How would you interpret the coefficient for seniority?

- The model with education has a lower residual sum of squares, thus reduces residuals more so than the model with seniority. We interpret the slope of seniority as for every one month increase of seniority income is expected to increase by .25. When seniority is 0, the expected mean of income is 39.

6. Examine `Income` as a function of `Education` and `Seniority` in linear regression

```
mod_edu_sen <- lm(Income ~ Education + Seniority, data=dat)
summary(mod_edu_sen)
```

```
##
## Call:
## lm(formula = Income ~ Education + Seniority, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.113 -5.718 -1.095  3.134 17.235
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -50.08564    5.99878  -8.349 5.85e-09 ***
## Education     5.89556    0.35703  16.513 1.23e-15 ***
## Seniority     0.17286    0.02442   7.079 1.30e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.187 on 27 degrees of freedom
## Multiple R-squared:  0.9341, Adjusted R-squared:  0.9292
## F-statistic: 191.4 on 2 and 27 DF,  p-value: < 2.2e-16
```

A. How would you put this into a formula?

$\overline{Income_i} = \beta_0 + \beta_1 \times Education_i + \beta_2 \times Seniority_i + \epsilon_i$

7. Do #2-6 seem to make sense?

- Models with education and seniority makes sense since they have a positive relationship with income. However, having both in the model also makes sense, yet in our sample we do not have have anyone with 0 education and 0 seniority realistically.

8. Can you use results from #2-6? Why and why not?

- We may need to center both of these variables and remodel again. We cannot infer on a causal outcome.