Week 8 Exercise

Calum Webb

13/11/2021

How large is the gender pay gap in the UK?

In this week's practical example, we will follow researchers using real data from the Labour Force Survey 2018 to explore the gender pay gap in the UK.

The Labour Force Survey is a repeated cross-sectional survey that includes data from a stratified sample of respondents and aims for these to be representative of the general population of working-age adults in the UK. The LFS' methodology at the time only allowed respondents to answer either 'male' or 'female', or allowed them to refuse to answer, and did not differentiate between sex and gender in its survey. Participants with 'missing' responses, including those who chose not to answer some questions, are not included in this teaching dataset extract. Given the mode of questionnaire delivery, where respondents are asked their gender or where interviewers record the participant's gender as the gender they felt they present as (yes, really, this happens!), responses should be considered as participant's gender rather than their sex assigned at birth. This limits our analysis of the real data to identifying any gender pay gap between people whose gender aligns to either male or female, and prevents us from exploring how pay may differ between these groups and people who may have a gender identity other than male or female.

We start by reading in the data and loading required packages. Remember to check the variable names and descriptions in the `codebook.xlsx` file.

```
library(car)
```

```
## Loading required package: carData
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
lfs_data <- read_csv("lfs_sample.csv")
```

```
## Rows: 8708 Columns: 53
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (31): ten1, housex, sexx, ages, ntnlty12, regionx, ethukeul, fbx, marsta...
## dbl (14): age, numchild04, numchild516, ayfl19, arrivalx, grsswk, hourpay, c...
## lgl  (8): ytetjb, ownbus, relbus, look4, lkyt4, wait, likewk, nolwm
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

The researchers start by creating a bivariate linear regression model to find out what the gender pay gap is without taking account of any other possible indirect pathways or confounding variables.

```
model_1 <- lm(data = lfs_data,
              formula = grsswk ~ sexx)

summary(model_1)
```

```
##
## Call:
## lm(formula = grsswk ~ sexx, data = lfs_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -495.41 -196.71  -57.41  149.29 1034.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  427.711      3.997  107.00   <2e-16 ***
## sexxmale     167.702      5.825   28.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 271.3 on 8706 degrees of freedom
## Multiple R-squared:  0.08694,    Adjusted R-squared:  0.08683
## F-statistic: 828.9 on 1 and 8706 DF,  p-value: < 2.2e-16
```

- Describe the size of the gender pay gap based on this model and whether it was statistically significant or not.

The model suggests that men in the Labour Force Survey were paid, on average, £167.70 higher than women in the sample per week. This result was statistically significant ($p < 0.05$). Gender was able to explain approximately 8.7% of the variance in gross weekly pay.

The researchers want to change the variables they are using so that men are the reference category, and women are the category coded as "1".

```
lfs_data <- lfs_data %>%
  mutate(
    female = case_when(is.na(sexx) ~ NA_real_,
                       sexx == "female" ~ 1,
```

```
                        TRUE ~ 0)
  )
```

- Re-run the model with this new variable where men are coded 0 and women are coded as 1. Save it as `model_2` and create a summary of the output. Describe the results.

```
model_2 <- lm(data = lfs_data,
              formula = grsswk ~ female)

summary(model_2)
```

```
##
## Call:
## lm(formula = grsswk ~ female, data = lfs_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -495.41 -196.71  -57.41  149.29 1034.29
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  595.413      4.237  140.54   <2e-16 ***
## female      -167.702      5.825  -28.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 271.3 on 8706 degrees of freedom
## Multiple R-squared:  0.08694,    Adjusted R-squared:  0.08683
## F-statistic: 828.9 on 1 and 8706 DF,  p-value: < 2.2e-16
```

This model shows that women in the sample were paid, on average, £167.70 less per week than men. This is the opposite of the above finding because the reference category has been changed from women to men. Everything else is identical because only the reference category has changed.

The researchers now add some additional independent variables to the model, hours worked per week (`ttushr`), number of children aged 0-4 (`numchild04`), and simplified highest qualification held (`levqul15`).

```
model_3 <- lm(data = lfs_data,
              formula = grsswk ~ female + numchild04 + ttushr + levqul15)

summary(model_3)
```

```
##
## Call:
## lm(formula = grsswk ~ female + numchild04 + ttushr + levqul15,
##     data = lfs_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -850.89 -139.65  -35.12   99.23 1001.04
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    34.8997    12.6814   2.752 0.005935 **
## female                        -91.6038     5.2570 -17.425  < 2e-16 ***
```

```
## numchild04                     -6.0398     4.9753  -1.214 0.224796
## ttushr                         11.5341     0.2365  48.778  < 2e-16 ***
## levqul15No qualifications     -49.6453    14.0289  -3.539 0.000404 ***
## levqul15NQF Level 2            22.3693    10.0818   2.219 0.026528 *
## levqul15NQF Level 3            46.9242     9.6407   4.867 1.15e-06 ***
## levqul15NQF Level 4 and above 180.7924     8.6295  20.951  < 2e-16 ***
## levqul15Other qualifications   -4.4366    13.1944  -0.336 0.736692
## levqul15Trade apprenticeships   2.1193    16.8420   0.126 0.899867
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 224.9 on 8597 degrees of freedom
##   (101 observations deleted due to missingness)
## Multiple R-squared:  0.3736, Adjusted R-squared:  0.3729
## F-statistic: 569.7 on 9 and 8597 DF,  p-value: < 2.2e-16
```

- How does the estimate for the `female` variable change now that these additional variables are included?

With the addition of the number of children in the household, the total hours worked per week, and the highest qualification achieved, the difference between women and men falls from -£167.70 to -£91.60. This suggests that some of the difference between the two groups could be 'explained' or mediated by differences in the numbers of dependent children in the household, the numbers of hours worked per week, and the highest level of qualification a person has. However, the new estimate remains statistically significant ($p < 0.05$).

- What value has been chosen as the reference category for the levqul15 variable? Hint: use `tabyl()` to see a summary of the number of people in each category.

```
lfs_data %>% tabyl(levqul15)
```

```
##               levqul15    n      percent valid_percent
##      Below NQF Level 2  843 0.0968075333    0.09689655
##      No qualifications  381 0.0437528709    0.04379310
##            NQF Level 2 1267 0.1454983923    0.14563218
##            NQF Level 3 1590 0.1825907212    0.18275862
##  NQF Level 4 and above 3929 0.4511943041    0.45160920
##   Other qualifications  456 0.0523656408    0.05241379
##  Trade apprenticeships  234 0.0268718420    0.02689655
##                   <NA>    8 0.0009186955            NA
```

The reference category chose by R was "Below NQF Level 2", as it is the first unique value in alphabetical order.

- Do you think the researchers should change their reference category for highest qualification? If so, which category do you think they should choose and why?

Yes, the results may be easier to interpret if the reference category were changed to either the most common category (NQF Level 4 and above) or a value at the lowest point of the scale (No qualifications). Either of these would be good choices for aiding interpretation.

The researchers decide to recode their `levqul15` variable so that "NQF Level 4 and above" (university-level education) is the reference category. They use the `factor()` and `relevel()` functions to do this.

```
lfs_data <- lfs_data %>%
  mutate(
    levqul15_fct = relevel(factor(levqul15), ref = "NQF Level 4 and above")
  )
```

- Write the code to re-run the multiple regression model with the new `levqul15_fct` variable created above. Save the new model as `model_4` and print a summary using `summary()`.

```
model_4 <- lm(data = lfs_data,
              formula = grsswk ~ female + numchild04 + ttushr + levqul15_fct)

summary(model_4)
```

```
##
## Call:
## lm(formula = grsswk ~ female + numchild04 + ttushr + levqul15_fct,
##     data = lfs_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -850.89 -139.65  -35.12   99.23 1001.04
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       215.6921    11.1975  19.263   <2e-16 ***
## female                            -91.6038     5.2570 -17.425   <2e-16 ***
## numchild04                         -6.0398     4.9753  -1.214    0.225
## ttushr                             11.5341     0.2365  48.778   <2e-16 ***
## levqul15_fctBelow NQF Level 2    -180.7924     8.6295 -20.951   <2e-16 ***
## levqul15_fctNo qualifications    -230.4377    12.2632 -18.791   <2e-16 ***
## levqul15_fctNQF Level 2          -158.4231     7.3376 -21.591   <2e-16 ***
## levqul15_fctNQF Level 3          -133.8683     6.7317 -19.886   <2e-16 ***
## levqul15_fctOther qualifications -185.2290    11.2630 -16.446   <2e-16 ***
## levqul15_fctTrade apprenticeships -178.6732    15.4111 -11.594   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 224.9 on 8597 degrees of freedom
##   (101 observations deleted due to missingness)
## Multiple R-squared:  0.3736, Adjusted R-squared:  0.3729
## F-statistic: 569.7 on 9 and 8597 DF,  p-value: < 2.2e-16
```

- Interpret the effect of each different kind of qualification on weekly pay relative to the new reference category of university level (Level 4) qualifications. E.g. "Respondents with no qualifications had weekly pay that was on average £X lower/higher than respondents with university level qualifications..."

After controlling for gender, the number of dependent children, and the total hours worked each week, respondents with no qualifications had weekly pay that was on average £230.44 lower than respondents with Level 4 (university) level qualifications ($p<0.05$). Respondents with qualifications below NQF level 2 had weekly pay that was on average £180.79 less; those with qualifications at NQF level 2 had weekly pay £158.42 lower ($p<0.05$); those with qualifications at NQF level 3 had weekly pay £133.87 lower ($p<0.05$); those with 'other' qualifications had weekly pay £185.23 lower ($p<0.05$), and those with trade apprenticeship qualifications had weekly pay £178.67 lower than those with level 4 (university or higher) qualifications ($p<0.05$).

To visualise the differences between educational group, while controlling for gender, number of dependent children, and hours worked, the researchers use the `ggeffects` package to create a prediction plot:

```
library(effects)
```

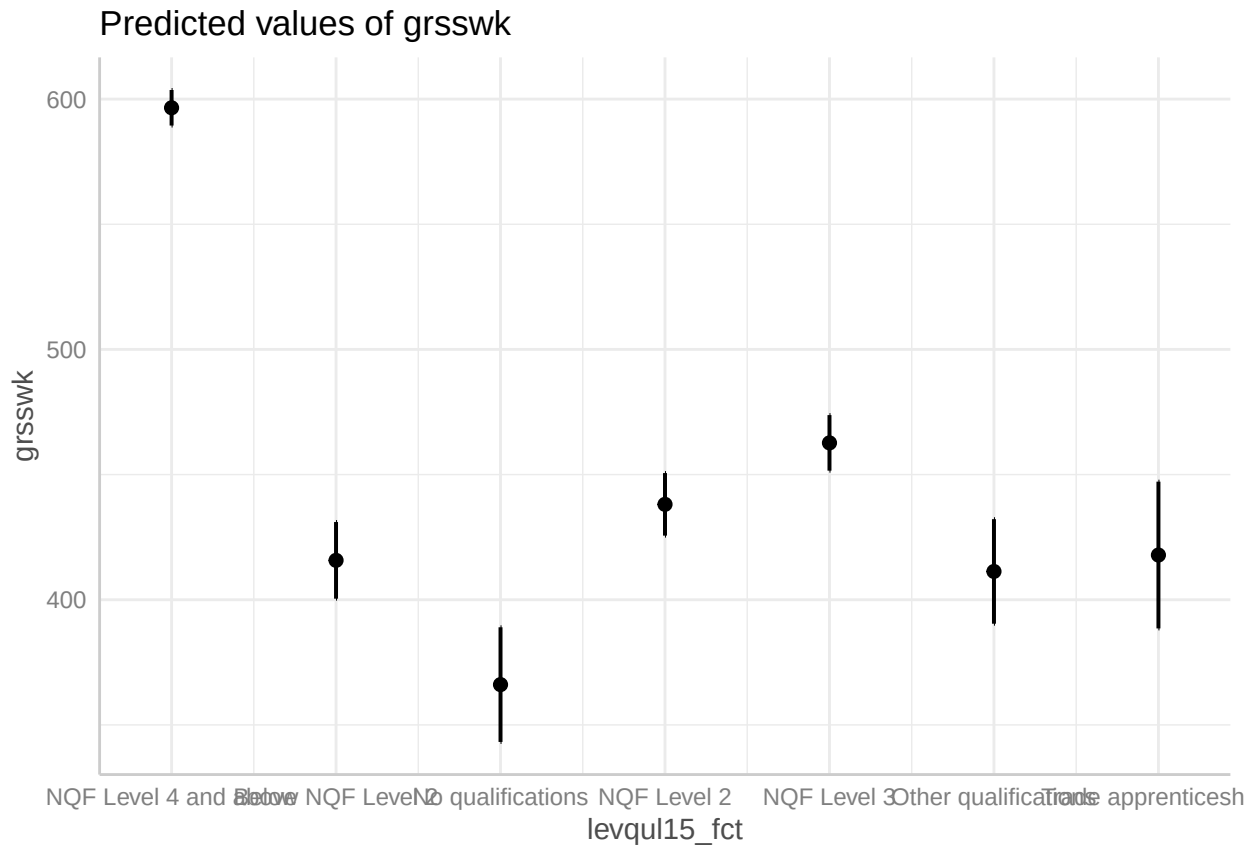
```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
library(ggeffects)

# ggeffect will create predictions at mean values of the other variables
# and varying values of the chosen variable (here: education);
# The plot() function then plots this as a forest plot (though if you)
# leave it out it will just give you the values in text form
ggeffect(model_4, terms = "levqul15_fct")
```

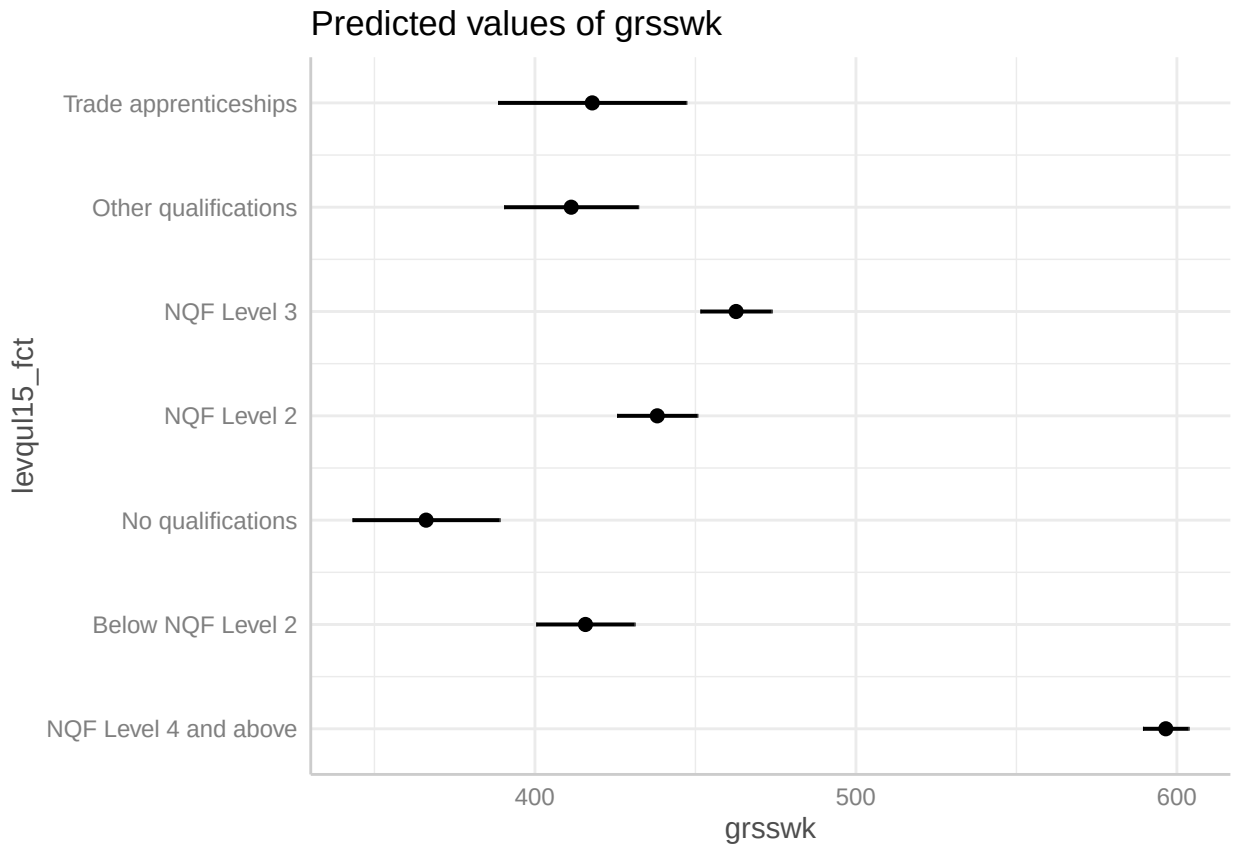
```
## # Predicted values of grsswk
##
## levqul15_fct          | Predicted |           95% CI
## ------------------------------------------------------
## NQF Level 4 and above |    596.53 | [589.43, 603.63]
## Below NQF Level 2     |    415.74 | [400.43, 431.05]
## No qualifications     |    366.09 | [343.19, 389.00]
## NQF Level 2           |    438.11 | [425.61, 450.60]
## NQF Level 3           |    462.66 | [451.56, 473.76]
## Other qualifications  |    411.30 | [390.42, 432.18]
## Trade apprenticeships |    417.86 | [388.54, 447.17]
```

```
ggeffect(model_4, terms = "levqul15_fct") %>% plot()
```



- Add the following function to the end of the ggeffect(model_4, terms = "levqul15_fct") %>% plot() code and check what changes: + coord_flip(). How does this change help the communication of the plot?

```
ggeffect(model_4, terms = "levqul15_fct") %>% plot() + coord_flip()
```



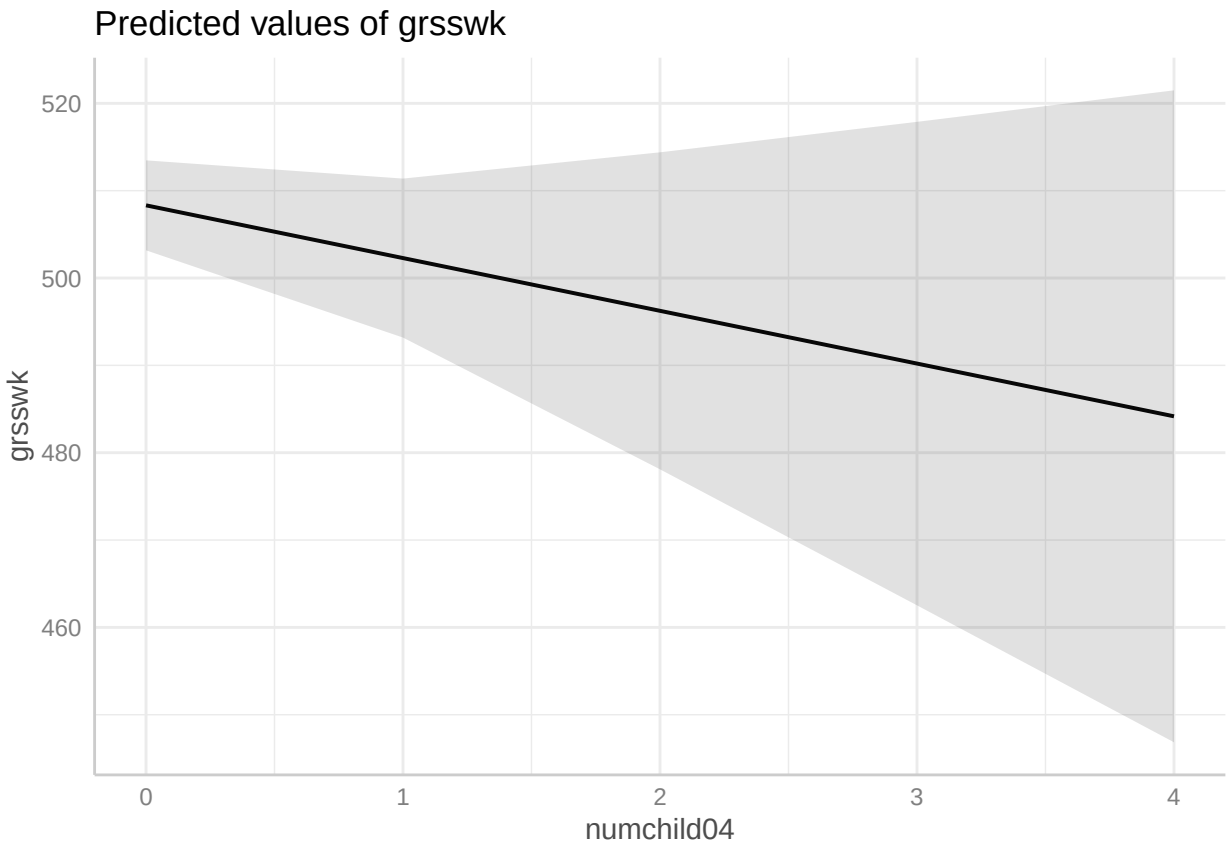
The change makes it easier to read the labels for each education level.

- What was the mean predicted income for people with NQF level 4 and above qualifications?

£596.53

- Now use the ggeffect code as a template to get the predicted pay for different numbers of dependent children by changing the terms argument to refer to the `numchild04` variable.

```
ggeffect(model_4, terms = "numchild04") %>% plot()
```



- What happens to the standard error (the shaded area) around our predictions for different numbers of dependent children as the number of dependent children increases? Why does this happen and how does it relate to the p-value associated with the numchild04 variable in the model?

As the number of dependent children increases, the error around the predictions gets wider. This is because there are a smaller number of people with 2 to 4 dependent children all aged between 0-4 in the sample than there are people with no dependent children that age, or with one child that age. The fact that the standard errors are so wide across the predictions is also reflected in the fact that the p-value is greater than 0.05 - we cannot be very certain that the mean income for people with more dependent children aged 0-4 is significantly lower after controlling for gender, education, and number of hours worked.

Lastly, the researchers decide to add whether the respondent is a manager, foreman/supervisor, or neither a manager or supervisor as a predictor of gross weekly pay in the model, using the `manager` variable.

```
model_5 <- lm(data = lfs_data,
              formula = grsswk ~ female + numchild04 + ttushr + levqul15_fct + manager)

summary(model_5)
```

```
##
## Call:
## lm(formula = grsswk ~ female + numchild04 + ttushr + levqul15_fct +
##     manager, data = lfs_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -897.14 -127.36  -27.90   90.69  967.43
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       249.6640    12.8641  19.408  < 2e-16 ***
## female                            -89.4297     4.9971 -17.896  < 2e-16 ***
## numchild04                        -10.1488     4.7305  -2.145   0.0319 *
## ttushr                              9.9242     0.2316  42.845  < 2e-16 ***
## levqul15_fctBelow NQF Level 2    -148.2600     8.2691 -17.929  < 2e-16 ***
## levqul15_fctNo qualifications    -189.9343    11.7410 -16.177  < 2e-16 ***
## levqul15_fctNQF Level 2          -136.0592     7.0152 -19.395  < 2e-16 ***
## levqul15_fctNQF Level 3          -117.9208     6.4198 -18.368  < 2e-16 ***
## levqul15_fctOther qualifications -147.1634    10.7785 -13.653  < 2e-16 ***
## levqul15_fctTrade apprenticeships -145.3801    14.6831  -9.901  < 2e-16 ***
## managerManager                    141.2225     8.3074  17.000  < 2e-16 ***
## managerNot manager or supervisor  -35.0647     7.4852  -4.685 2.85e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 213.7 on 8591 degrees of freedom
##   (105 observations deleted due to missingness)
## Multiple R-squared:  0.4348, Adjusted R-squared:  0.4341
## F-statistic: 600.8 on 11 and 8591 DF,  p-value: < 2.2e-16
```

- What is the reference category chosen automatically by `R` for the manager variable (Hint: Check using a `tabyl()`)

The reference category chosen by R for the `manager` variable was "Foreman or supervisor", because it is the first value when sorted alphabetically.

- Use the `relevel` and `factor` functions to create a new variable based on the `manager` variable that has "Not manager or supervisor" as the reference category.

```
lfs_data <- lfs_data %>%
  mutate(
    manager_fct = relevel(factor(manager), ref = "Not manager or supervisor")
  )
```

- Re-run the `model_5` multiple regression model but this time use the new manager variable that you created above, where the reference category is "Not manager or supervisor". Name this regression model `model_6`.

```
model_6 <- lm(data = lfs_data,
              formula = grsswk ~ female + numchild04 + ttushr + levqul15_fct + manager_fct)

summary(model_6)
```

```
##
## Call:
## lm(formula = grsswk ~ female + numchild04 + ttushr + levqul15_fct +
##     manager_fct, data = lfs_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -897.14 -127.36  -27.90   90.69  967.43
##
## Coefficients:
```

```
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        214.5993    10.6435  20.163  < 2e-16 ***
## female                             -89.4297     4.9971 -17.896  < 2e-16 ***
## numchild04                         -10.1488     4.7305  -2.145   0.0319 *
## ttushr                               9.9242     0.2316  42.845  < 2e-16 ***
## levqul15_fctBelow NQF Level 2     -148.2600     8.2691 -17.929  < 2e-16 ***
## levqul15_fctNo qualifications     -189.9343    11.7410 -16.177  < 2e-16 ***
## levqul15_fctNQF Level 2           -136.0592     7.0152 -19.395  < 2e-16 ***
## levqul15_fctNQF Level 3           -117.9208     6.4198 -18.368  < 2e-16 ***
## levqul15_fctOther qualifications  -147.1634    10.7785 -13.653  < 2e-16 ***
## levqul15_fctTrade apprenticeships -145.3801    14.6831  -9.901  < 2e-16 ***
## manager_fctForeman or supervisor    35.0647     7.4852   4.685 2.85e-06 ***
## manager_fctManager                 176.2872     5.7967  30.411  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 213.7 on 8591 degrees of freedom
##   (105 observations deleted due to missingness)
## Multiple R-squared:  0.4348, Adjusted R-squared:  0.4341
## F-statistic: 600.8 on 11 and 8591 DF,  p-value: < 2.2e-16
```

- Report the results from the model in full, describe the findings across all variables as well as the size and significance of the gender pay gap after controlling for all of these additional factors.

A multiple regression model predicting gross weekly pay was estimated with gender, number of children aged 0-4, hours worked per week, highest qualification, and managerial status as predictors. Overall, the model was able to explain approximately 43.5% of the variance in gross weekly pay.

Gender, number of dependent children aged 0-4, hours worked per week, all forms of qualifications below university-level qualifications (NQF level 4), and managerial or supervisor status were significantly associated with gross weekly pay ($p<0.05$). Women earned on average £89.43 less than men per week, after controlling for qualifications, number of children in household, hours worked, and managerial status. Each additional child in the household was associated with a reduction in weekly pay of £10.15. Each additional hour worked was associated with a £9.92 increase in weekly gross pay.

Qualifications below NQF4, or in other categories, were associated with lower earnings per week. Respondents with no qualifications earned on average £189.93 less per week; respondents with qualifications below Level 2 earned on average £148.26 less per week; respondents with NQF Level 2 earned on average £136.06 less per week; respondents with NQF level 3 earned £117.92 less on average per week. Respondents with other qualifications earned on average £147.16 less each week than those with Level 4 or higher qualifications; those with trade apprenticeship qualifications earned on average £145.38 per week, controlling for all other factors.

Lastly, those in foreman or supervisory roles earned on average £35.06 more each week than those not in supervisory or managerial positions. Respondents in managerial positions earned approximately £176.29 more per week than those with no managerial or supervisory responsibilities.

(Note that just this is around 300 words! You can see how your assessment words soon get depleted when you have multiple regression models)

- Are there any other variables in the codebook that may confound the relationship between gender and pay? If so, justify why you think the one you chose might be a potential confounder.

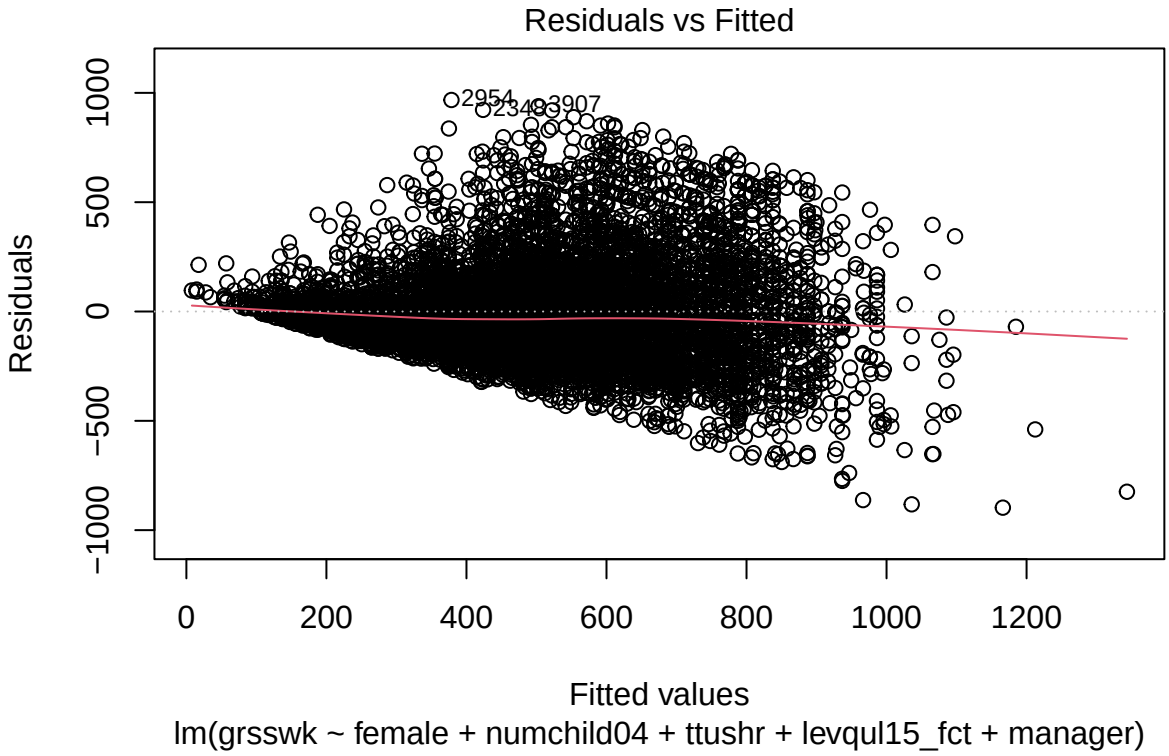
`sc10mmj` Occupational group is likely to be an important predictor of gross weekly pay, as some professions or areas of work are likely to be paid differently (e.g. financial sector versus care sector). Occupations are often also highly segregated by gender, which may be related to the gender pay gap.

Now, the researchers decide to assess whether their model violates any assumptions (linearity, heteroscedasticity, non-normality of residuals, outliers/leverage points, and multicollinearity).

The researchers decide to use a residuals versus fitted values plot to explore whether there was any evidence of non-normality in the model.

NB: They use model_5 as the results from these plots would be identical, even if the reference category has changed.

```
# Residuals Versus Fitted Values Plot
plot(model_5, which = 1)
```

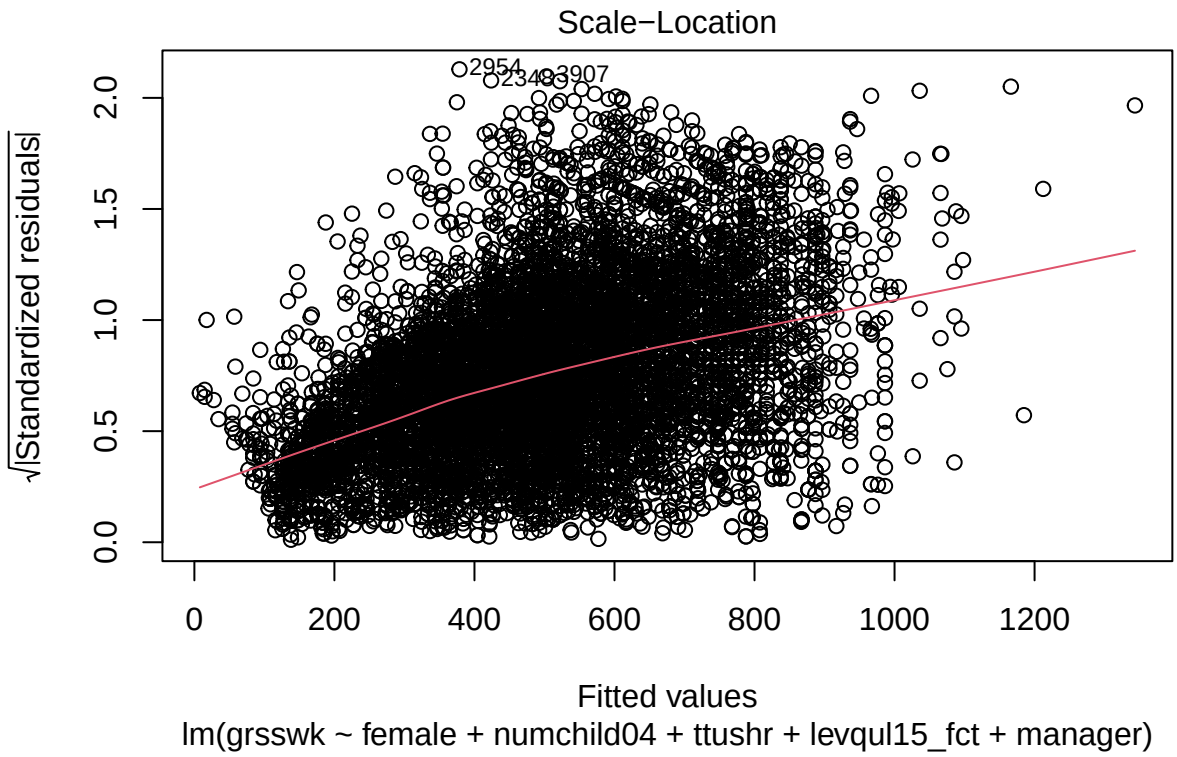


- Interpret the residuals versus fitted values plot (you can use the cheat sheet in the lecture slides). Was there any evidence of non-normality? If so, how severe does it look and what consequences might it have on the model's accuracy and bias?

The diagnostic plot suggests that a linear fit to the data was relatively good, as it remained relatively close to the 0 point. There is some evidence that a non-linear fit may be more appropriate for higher values of the independent variables.

Next, the researchers chose to inspect the homoscedasticity of the residuals using a scale-location plot.

```
# Scale-Location Plot
plot(model_5, which = 3)
```

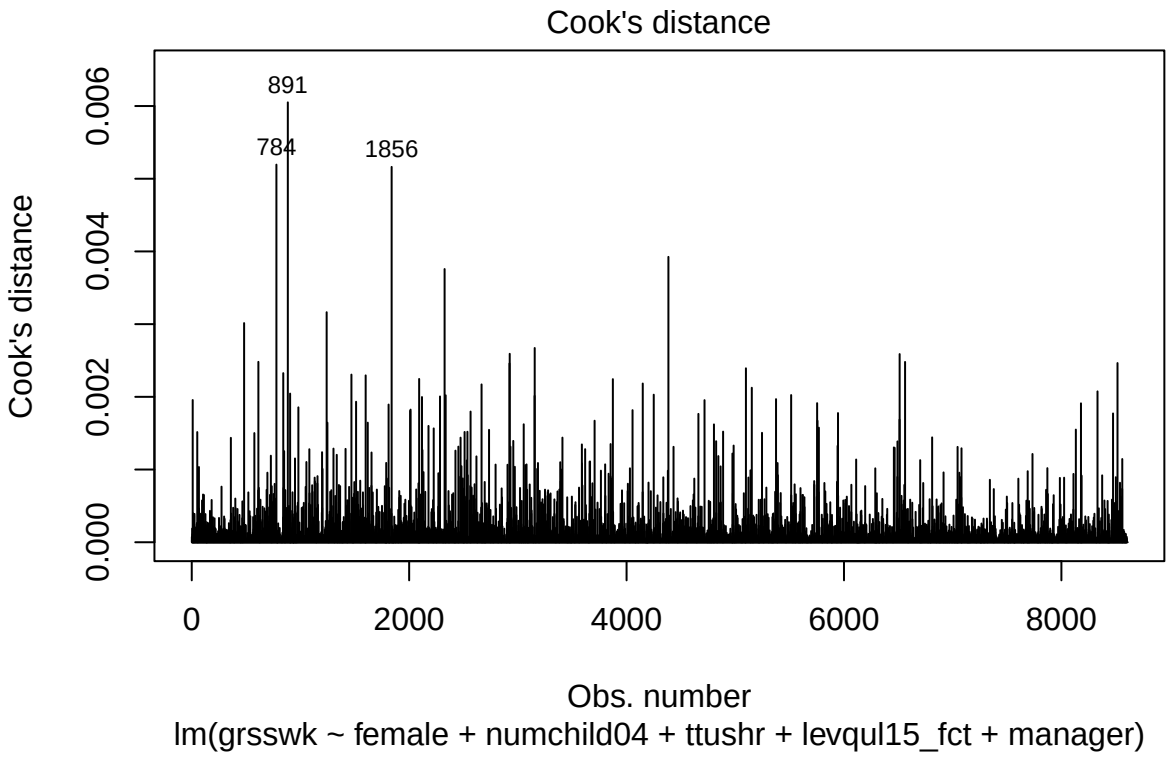


- Is there evidence of heteroscedasticity among the residuals? If so, what shape does this heteroscedasticity take and what affect might this have on the model estimates.

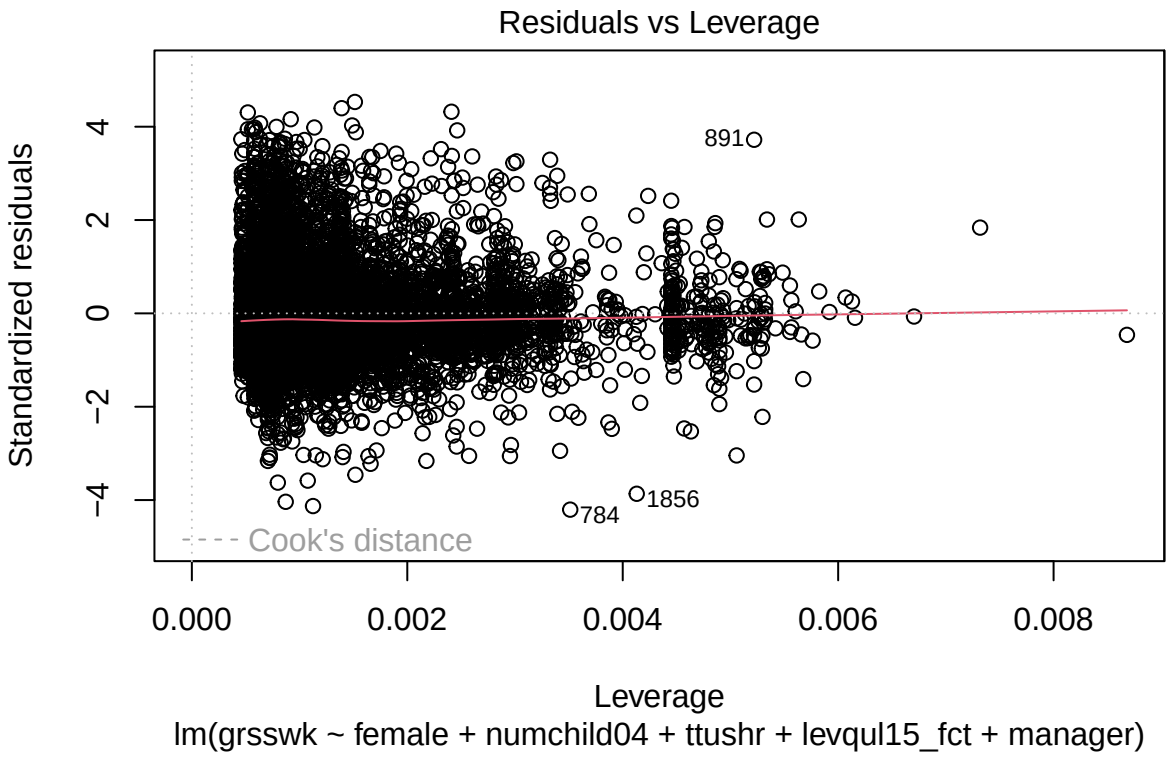
There appears to be quite strong evidence of heteroscedasticity in this data; as the fitted values tend to increase, the variance around the estimate also tends to increase. This indicates that some standard errors may be biased, affecting any marginal p-values. A regression model using a Weighted Least Squares estimator may be more appropriate to verify results.

The researchers then checked whether there was evidence of significant outliers in the model using Cook's distance plots.

```
# Cook's Distance Plot
plot(model_5, which = 4)
```




```
# Residuals versus leverage plot
plot(model_5, which = 5)
```

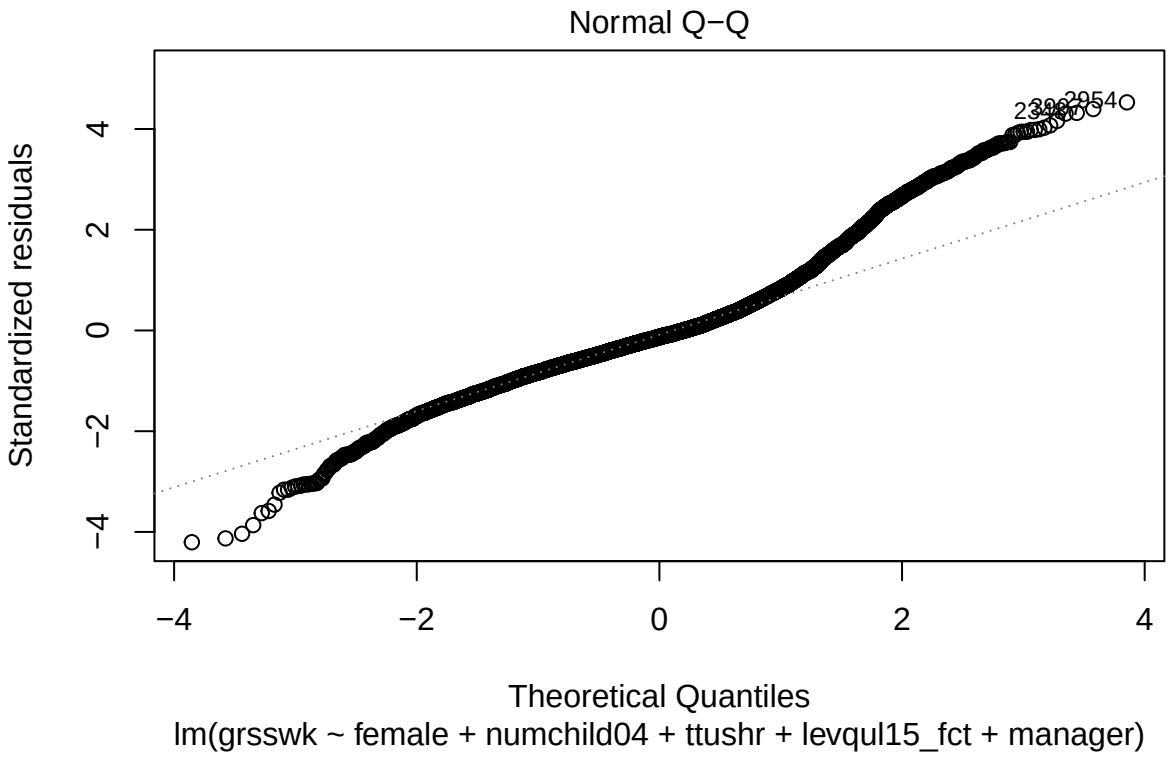


- Based on the above plots, was there any evidence of potentially significant outliers? If so, what is the observation row number of these outliers?

The Cook's distance plots suggest that observation 891, 784, and 1856 may be potentially influential outliers. It may be beneficial to re-run the analysis with these outliers removed to ensure that the estimates do not change substantially. However, the leverage plot suggests that the outliers are not substantially influencing the regression estimates (as the red line is relatively horizontal and close to 0).

The normality of the residuals were then checked using a Q-Q plot.

```
# Q-Q plot
plot(model_5, which = 2)
```



- Was there any evidence of non-normality of residuals based on the Q-Q plot? If so, what impact might this have on our model results (Hint: use the Cheat Sheet in the lecture slides for week 7 or week 8)

The Q-Q plot shows that there was some evidence of non-normality of residuals in the final model, as the points deviate consistently from the diagonal line. While existing research shows that this is unlikely to affect our estimates (due to the large sample size), our standard errors may be biased. However, due to the relatively small standard errors in many estimates, this is unlikely to substantially change our conclusions. However, a log-transformation of the dependent variable may improve the normality of residuals.

Finally, the researchers explore whether there may be any multicollinearity present among their independent variables using the `vif` function from the `car` package.

```
# Variance Inflation Factor
car::vif(model_5)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## female       1.171353  1        1.082290
## numchild04   1.012264  1        1.006113
## ttushr       1.232178  1        1.110035
## levqul15_fct 1.086590  6        1.006944
## manager      1.126594  2        1.030248
```

- Do the VIF/GVIF values suggest there is any degree of multicollinearity that could bias the model estimates? Why/why not?

The VIF statistics suggest that there is no significant concerns around multicollinearity in the model. All VIF statistics are close to 1, and relatively far from the values of 5 or higher that would indicate a potential

problem in the model.

Week 8 Challenges

- Try rerunning the last model in the script but with a logged (`log()`) or square-root (`sqrt()`) version of the dependent variable (`grsswk`). Does this change the model R-squared or any of the assumption checks? If so, how do these change and what would the possible consequences be.
- Try adding the variable you identified as an additional potential confounder to the model: how does adding this variable change the findings? Can you use a prediction plot to show the predicted means for each value of your chosen variable?
- Conduct a similar analysis to the above, but this time explore the pay gap between ethnic groups in the UK.