



How Does Health Vary Across Regions in the UK?

Multiple Linear Regression model with Qualitative Variables

Zi Ye

ENVS225

Exploring the Social World

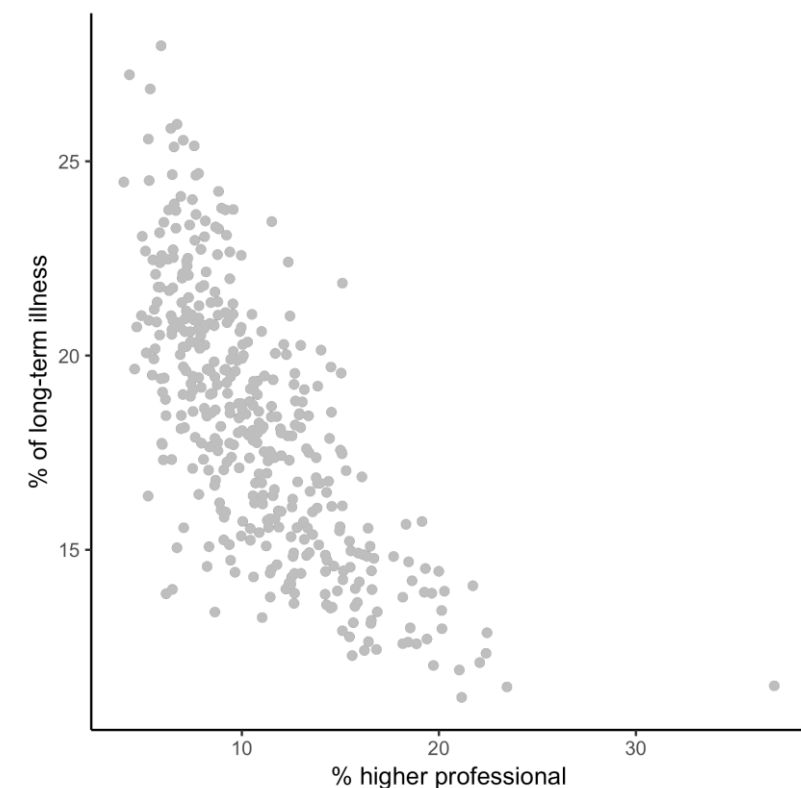
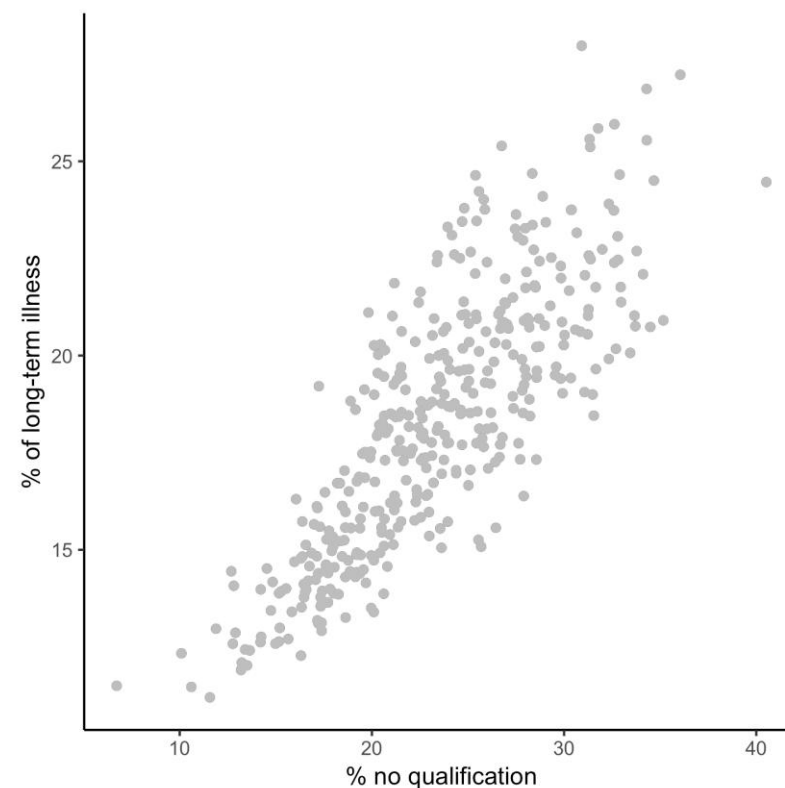


UNIVERSITY OF
LIVERPOOL

So far - Quantitative Attributes

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 X3 + \epsilon$$

Scale/Continuous variables



But - Qualitative variables

- Where is the rate of long-term illness higher in the UK? Does it vary much across regions?
- Is the rate of long-term illness higher in the North West?
- Other examples:
 - Is there a **gender** pay gap?
 - How does **ethnicity** influence the industry people work?
 - How does **marital status** influence house ownership?
 - What **age group** is more migratory?

While continuous variables capture quantitative effects, categorical variables provide insights into differences across groups.

Learning Outcomes

Aim: Understanding how to estimate and interpret a regression model using qualitative variables



What Are
Qualitative/Categorical
Variables?



How to interpret regression
model with qualitative
variable(s)

What Are Qualitative Variables?

Qualitative variables

Categorical variable

- Nominal: categorical data without natural order.
- Ordinal: categorical data with a meaningful order.
- Gender
- Location
- Marital status
- Ethnicity
- Grade
- Educational level
- Socio-economic status
- etc.

Working with Different Data Types

➤ Nominal



3 What is your sex?

➤ A question about gender identity will follow if you are aged 16 or over

☐ Female

☐ Male

Qualitative variable:
Gender

id	gender
1	female
2	female
3	male
4	female

2 Dummy variables

id	female	male
1	1	0
2	1	0
3	0	1
4	1	0

Working with Different Data Types

➤ Ordinal

25 Which of these qualifications do you have?

➤ Tick **every** box that applies if you have **any** of the qualifications listed

➤ If your UK qualification is not listed, tick the box that contains its nearest equivalent

➤ If you have qualifications gained outside the UK, tick the 'Foreign qualifications' box and the nearest UK equivalents (if known)

☐ 1 - 4 O levels / CSEs / GCSEs (any grades), Entry Level, Foundation Diploma

Qualitative variable:
Education

5 Dummy variables

id	qualification
1	degree
2	No qual
3	No qual
4	2+ A levels

id	Level 1 and entry	1 A level	2+ A levels	Degree	no qualification
1	0	0	0	1	0
2	0	0	0	0	1
3	0	0	0	0	1
4	0	0	1	0	0

Working with Different Data Types

Scale/Continuous data as categorical variables

Age -> Age band
Income -> Income band
Housing cost -> cost band

3 What is your date of birth?

Day

Month

Year

10 Dummy variables

id	age
1	20
2	45
3	70
4	35

id	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65+
1	1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	1	0	0	0	0
3	0	0	0	0	0	0	0	0	0	1
4	0	0	0	1	0	0	0	0	0	0

Wrap up !

1 Original Variables
with n Categories



n Dummy Variables

3 Dummy variables

id	Residence
1	NW
2	SE
3	L
4	NW

Residence region

id	North West	South East	London
1			0
2			0
3			1
4	1	0	0





Regression with Qualitative Variables: **Set the reference variable**

- For the regression model, include **one less dummy variable** (base or reference category) than the number of categories

Variable

id	gender
1	female
2	female
3	male
4	female

Dummy Variable

id	female	male
1	1	0
2	1	0
3	0	1
4	1	0

Regression model

id	female
1	1
2	1
3	0
4	1

OR

id	male
1	0
2	0
3	1
4	0

Wrap up again !!

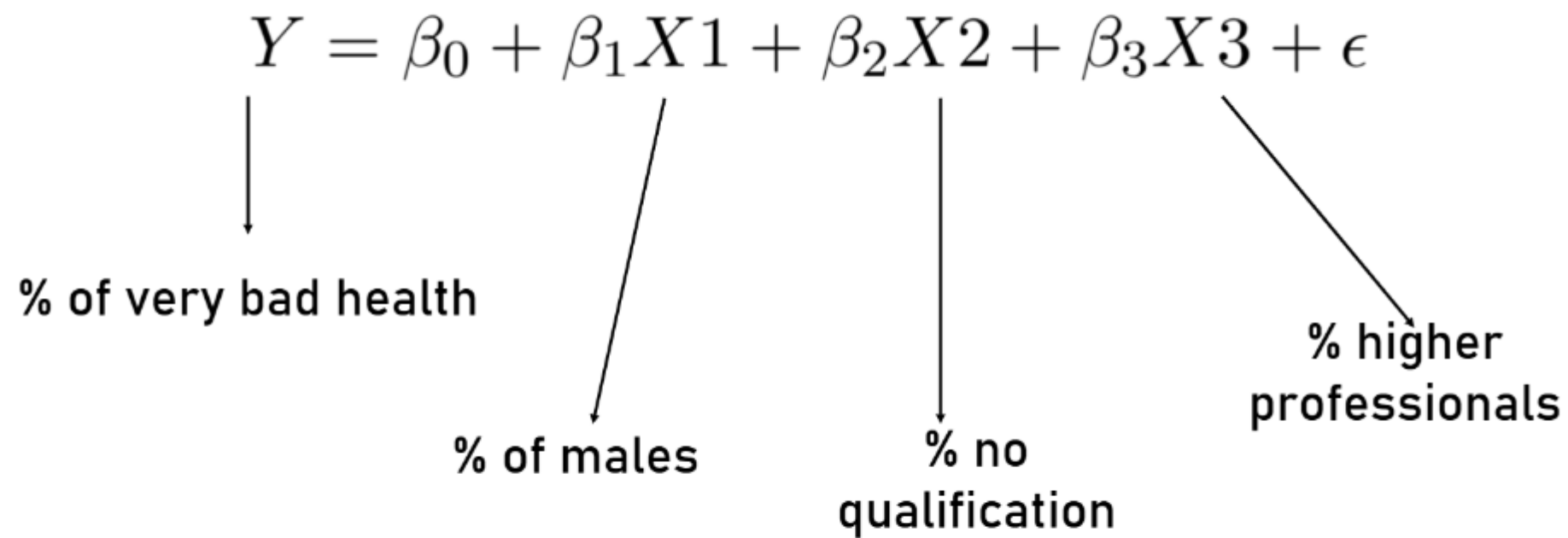


In Practical:

- `df$QualVar <- fct_relevel(df$QualVar, "reference_value")`
- `model <- lm(Y ~ X1 + X2 + X3 + .. + Xn + QualVar, data = df)`
- `summary(model)`

Interpretation

Last week



We are trying to find the β s

This week, with qualitative variables

qualitative variable **d** with $N=s$
different categories

% of long-term illness

The s dummy variables of **d**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{d1} X_{d1} + \beta_{d1} X_{d2} + \dots + \beta_{ds} X_{ds} + \varepsilon$$

% of males

% no
qualification

% higher
professionals

We are still trying to find all the β

Regression Model

- Y = % of long-term illness
- Original qualitative variable **Region** has 12 categories ($n=12$)
- X_1, X_2, \dots, X_{12} : Dummy variables for UK's region
- 11 of the dummy variables will be used in the model
- R will take care of creating dummy variables for you automatically! **BUT**, please let R know which one you want to set as **reference category**


```
df$Region_label <- factor(df$Region,c(1:12),labels=c("East Midlands",  
                                                    "East of England",  
                                                    "London",  
                                                    "North East",  
                                                    "North West",  
                                                    "South East",  
                                                    "South West",  
                                                    "West Midlands",  
                                                    "Yorkshire and the Humber",  
                                                    "Wales",  
                                                    "Scotland",  
                                                    "Northern Ireland"))
```

Therefore, first, we set London as the reference:

```
df$Region_label <- fct_relevel(df$Region_label, "London")
```

Similar to last week, we build our linear regression model, but also include the *Region_label* variable into the model.

```
model <- lm(pct_Long_term_ill ~ pct_Males + pct_No_qualifications + pct_Higher_manager_prof + Region_label, data =  
summary(model)
```

Relative to The Reference Category

Call:

```
lm(formula = pct_Long_term_ill ~ pct_Males + pct_No_qualifications +  
    pct_Higher_manager_prof + Region_label, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.2963	-0.9090	-0.1266	0.8168	5.2821

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	41.54134	5.22181	7.955	1.95e-14	***
pct_Males	-0.75756	0.10094	-7.505	4.18e-13	***
pct_No_qualifications	0.50573	0.03062	16.515	< 2e-16	***
pct_Higher_manager_prof	0.08910	0.03674	2.426	0.01574	*
Region_labelEast Midlands	1.14167	0.35015	3.260	0.00121	**
Region_labelEast of England	-0.01113	0.33140	-0.034	0.97322	
Region_labelNorth East	2.70447	0.49879	5.422	1.03e-07	***
Region_labelNorth West	2.64240	0.35468	7.450	6.03e-13	***
Region_labelSouth East	0.48327	0.30181	1.601	0.11013	
Region_labelSouth West	2.62729	0.34572	7.600	2.22e-13	***
Region_labelWest Midlands	0.91064	0.37958	2.399	0.01690	*
Region_labelYorkshire and the Humber	1.03930	0.41050	2.532	0.01174	*
Region_labelWales	4.63424	0.41368	11.202	< 2e-16	***
Region_labelScotland	0.46291	0.38916	1.189	0.23497	
Region_labelNorthern Ireland	0.55722	0.42215	1.320	0.18762	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.394 on 391 degrees of freedom

Multiple R-squared: 0.8298, Adjusted R-squared: 0.8237

F-statistic: 136.218 on 14 and 391 DF, p-value: < 2.2e-16

What variable is the
reference category?

London

Can you Compare Dummies?

Call:

```
lm(formula = pct_Long_term_ill ~ pct_Males + pct_No_qualifications +  
    pct_Higher_manager_prof + Region_label, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.2963	-0.9090	-0.1266	0.8168	5.2821

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	41.54134	5.22181	7.955	1.95e-14	***
pct_Males	-0.75756	0.10094	-7.505	4.18e-13	***
pct_No_qualifications	0.50573	0.03062	16.515	< 2e-16	***
pct_Higher_manager_prof	0.08910	0.03674	2.426	0.01574	*
Region_labelEast Midlands	1.14167	0.35015	3.260	0.00121	**
Region_labelEast of England	-0.01113	0.33140	-0.034	0.97322	
Region_labelNorth East	2.70447	0.49879	5.422	1.03e-07	***
Region_labelNorth West	2.64240	0.35468	7.450	6.03e-13	***
Region_labelSouth East	0.48327	0.30181	1.601	0.11013	
Region_labelSouth West	2.62729	0.34572	7.600	2.22e-13	***
Region_labelWest Midlands	0.91064	0.37958	2.399	0.01690	*
Region_labelYorkshire and the Humber	1.02000	0.37958	2.689	0.00800	**



CAN'T say the estimated percentage of long-term ill population in the North West is lower than in the North East!

Residual standard error: 1.394 on 391 degrees of freedom

Multiple R-squared: 0.8298, Adjusted R-squared: 0.8237

F-statistic: 136.2 on 14 and 391 DF, p-value: < 2.2e-16

But You Can Change The Base Category and re-run the model

```
df$Region_label <- fct_relevel(df$Region_label, "North East")
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	44.24582	5.20125	8.507	3.85e-16	***
pct_Males	-0.75756	0.10094	-7.505	4.18e-13	***
pct_No_qualifications	0.50573	0.03062	16.515	< 2e-16	***
pct_Higher_manager_prof	0.08910	0.03674	2.426	0.015738	*
Region_labelLondon	-2.70447	0.49879	-5.422	1.03e-07	***
Region_labelEast Midlands	-1.56281	0.46292	-3.376	0.000809	***
Region_labelEast of England	-2.71561	0.45836	-5.925	6.87e-09	***
Region_labelNorth West	-0.06208	0.46209	-0.134	0.893206	
Region_labelSouth East	-2.22120	0.45667	-4.864	1.67e-06	***
Region_labelSouth West	-0.07718	0.47482	-0.163	0.870957	
Region_labelWest Midlands	-1.79384	0.48230	-3.719	0.000229	***
Region_labelYorkshire and the Humber	-1.66517	0.50695	-3.285	0.001113	**
Region_labelWales	1.92976	0.50111	3.851	0.000137	***
Region_labelScotland	-2.24157	0.47299	-4.739	3.01e-06	***
Region_labelNorthern Ireland	-2.14725	0.49296	-4.356	1.70e-05	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.394 on 391 degrees of freedom

Multiple R-squared: 0.8298, Adjusted R-squared: 0.8237

F-statistic: 136.2 on 14 and 391 DF, p-value: < 2.2e-16

What variable is the reference category now?

Is the model robust for prediction?

Call:

```
lm(formula = pct_Long_term_ill ~ pct_Males + pct_No_qualifications +  
    pct_Higher_manager_prof + Region_label, data = df)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.2963	-0.9090	-0.1266	0.8168	5.2821

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	41.54134	5.22181	7.955	1.95e-14	***
pct_Males	-0.75756	0.10094	-7.505	4.18e-13	***
pct_No_qualifications	0.50573	0.03062	16.515	< 2e-16	***
pct_Higher_manager_prof	0.08910	0.03674	2.426	0.01574	*
Region_labelEast Midlands	1.14167	0.35015	3.260	0.00121	**
Region_labelEast of England	-0.01113	0.33140	-0.034	0.97322	
Region_labelNorth East	2.70447	0.49879	5.422	1.03e-07	***
Region_labelNorth West	2.64240	0.35468	7.450	6.03e-13	***
Region_labelSouth East	0.48327	0.30181	1.601	0.11013	
Region_labelSouth West	2.62729	0.34572	7.600	2.22e-13	***
Region_labelWest Midlands	0.91064	0.37958	2.399	0.01690	*
Region_labelYorkshire and the Humber	1.03930	0.41050	2.532	0.01174	*
Region_labelWales	4.63424	0.41368	11.202	< 2e-16	***
Region_labelScotland	0.46291	0.38916	1.189	0.23497	
Region_labelNorthern Ireland	0.55722	0.42215	1.320	0.18762	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.394 on 391 degrees of freedom

Multiple R-squared: 0.8298, Adjusted R-squared: 0.8237

F-statistic: 136.2 on 14 and 391 DF, p-value: < 2.2e-16

Predicting

Using Estimates

What is the % of long-term illness in the NW, if the % Male is 49.8%, the % no qualification is 23.3% and the % of higher manager prof is 11.2%?

% long-term illness =

$$41.541 - 0.758 * (\% \text{ Male}) + 0.506 * (\% \text{ no qualification}) + 0.089 * (\% \text{ higher manager prof}) + 1.142 * \text{EM} + 2.704 * \text{NE} + 2.642 * \text{NW} + 2.627 * \text{SW} + 0.911 * \text{WM} + 1.039 * \text{YH} + 4.634 * \text{Wales}$$

What is the estimated % long-term illness for the North West?

$$41.541 - 0.758 * (49.8) + 0.506 * (23.3) + 0.089 * (11.2) + 1.142 * 0 + 2.704 * 0 + 2.642 * 1 + 2.627 * 0 + 0.911 * 0 + 1.039 * 0 + 4.634 * 0$$

= 19.22

Using Estimates

What is the estimated % long-term illness for the North East, with all the % are the same?

$$\begin{aligned} &41.541 - 0.758*(49.8) + 0.506*(23.3) + 0.089*(11.2) \\ &+ 1.142*0 + 2.704*1 + 2.642*0 + 2.627*0 + 0.911*0 + \\ &1.039*0 + 4.634*0 \\ &= 19.28 \end{aligned}$$

If we want to estimate the result for London, think about what the equation will be?

$$\begin{aligned} &41.541 - 0.758*(\%) + 0.506*(\%) + 0.089*(\%) \\ &+ 1.142*0 + 2.704*0 + 2.642*0 + 2.627*0 + 0.911*0 + \\ &1.039*0 + 4.634*0 \end{aligned}$$

But we can use R to do the estimation/prediction directly

```
obj_London <- data.frame(  
  pct_Males = 49.7,  
  pct_No_qualifications = 24.3,  
  pct_Higher_manager_prof = 14.7,  
  Region_label = "London"  
)  
obj_NW <- data.frame(  
  pct_Males = 49.8,  
  pct_No_qualifications = 23.3,  
  pct_Higher_manager_prof = 11.2,  
  Region_label = "North West"  
)  
obj_NE <- data.frame(  
  pct_Males = 49.8,  
  pct_No_qualifications = 23.3,  
  pct_Higher_manager_prof = 11.2,  
  Region_label = "North East"  
)
```

```
predict(model1, obj_London)
```

```
predict(model1, obj_NW)
```

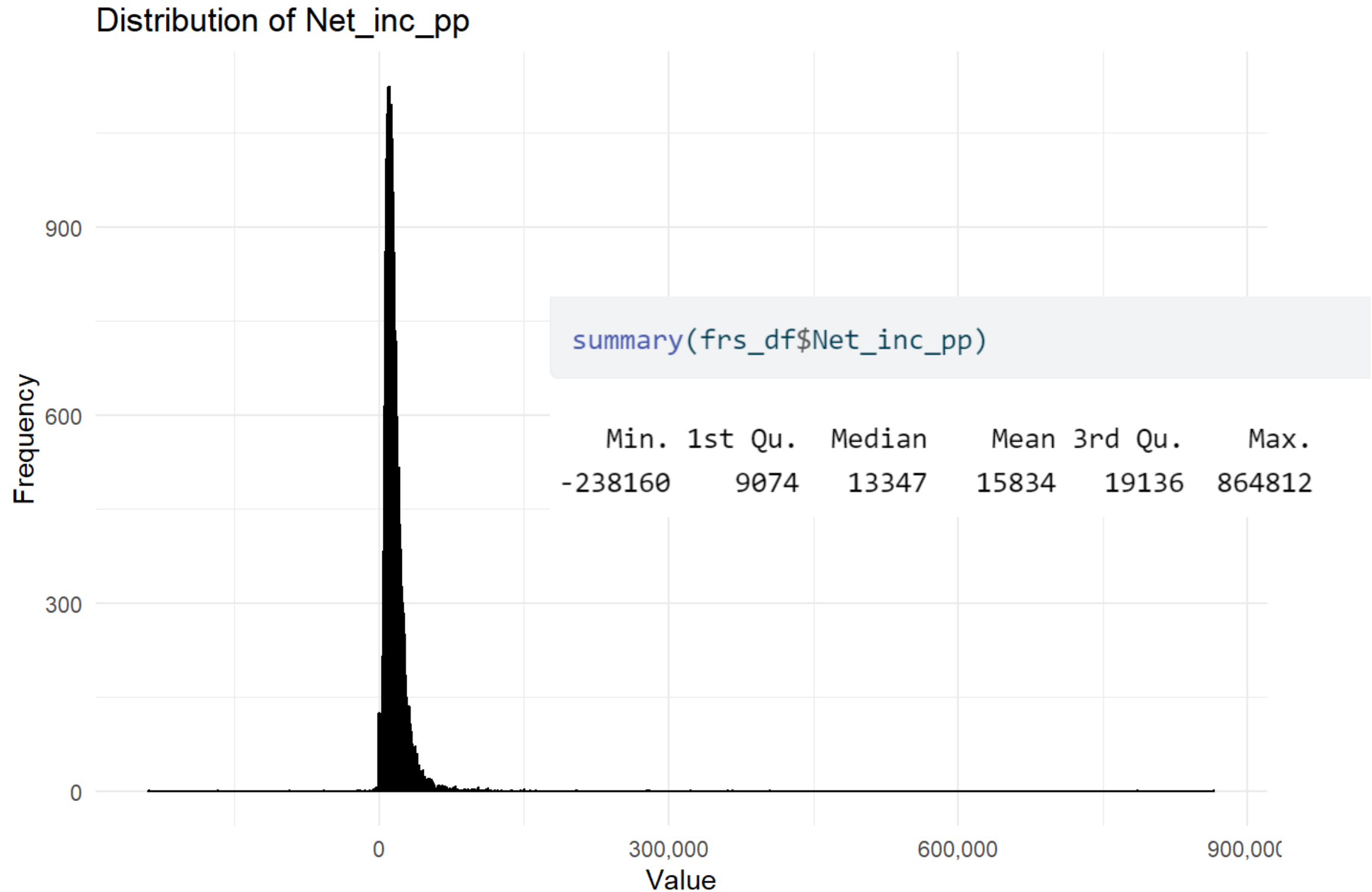
```
predict(model1, obj_NE)
```

How large is the gender gap in
the UK?

Data Prep. & Regression Model

- $Y = \text{Per capita household income}$
 - filter: household representative
 - income / family size
- X s: Qualitative variables for gender & general health
 - Original categories
 - Gender: Male & Female
 - Health: Very Bad, Bad, Fair, Good, Very Good

Net Household Income Per Capita Distribution



*What is the Expected Relationship Between **Income** & **Male/Good Health**?*

Positive/Positive?

Negative/Positive?

Negative/Negative?

```
frs_df$sex <- fct_relevel(as.factor(frs_df$sex), "Female")  
frs_df$health <- fct_relevel(as.factor(frs_df$health), "Very Bad")
```

Implement the regression model with the two qualitative independent variables.

```
model_frs <- lm(Net_inc_pp ~ sex + health, data = frs_df)  
summary(model_frs)
```

Is there a Gender Gap?

How Large is This?

Call:

```
lm(formula = Net_inc_pp ~ sex + health, data = frs_df)
```

Residuals:

Min	1Q	Median	3Q	Max
-255133	-6547	-2213	3515	845673

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	12115.5	762.9	15.881	< 2e-16	***
sexMale	2091.2	240.6	8.691	< 2e-16	***
healthBad	-102.8	854.3	-0.120	0.904205	
healthFair	1051.3	789.0	1.332	0.182751	
healthGood	2766.0	777.4	3.558	0.000375	***
healthVery Good	4931.8	787.8	6.260	3.95e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15530 on 16821 degrees of freedom

Multiple R-squared: 0.01646, Adjusted R-squared: 0.01616

F-statistic: 56.29 on 5 and 16821 DF, p-value: < 2.2e-16

**Baseline Sex:
Female**

Is There a Health Gradient?

Does Health Affects Individual Salary?

Call:

```
lm(formula = Net_inc_pp ~ sex + health, data = frs_df)
```

Residuals:

Min	1Q	Median	3Q	Max
-255133	-6547	-2213	3515	845673

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	12115.5	762.9	15.881	< 2e-16 ***
sexMale	2091.2	240.6	8.691	< 2e-16 ***
healthBad	-102.8	854.3	-0.120	0.904205
healthFair	1051.3	789.0	1.332	0.182751
healthGood	2766.0	777.4	3.558	0.000375 ***
healthVery Good	4931.8	787.8	6.260	3.95e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15530 on 16821 degrees of freedom

Multiple R-squared: 0.01646, Adjusted R-squared: 0.01616

F-statistic: 56.29 on 5 and 16821 DF, p-value: < 2.2e-16

**Baseline Health:
Very Bad**

Is the model good enough to do predict/estimate Net Household Income Per Capita by one's Health situation and Gender?

Call:

```
lm(formula = Net_inc_pp ~ sex + health, data = frs_df)
```

Residuals:

Min	1Q	Median	3Q	Max
-255133	-6547	-2213	3515	845673

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	12115.5	762.9	15.881	< 2e-16	***
sexMale	2091.2	240.6	8.691	< 2e-16	***
healthBad	-102.8	854.3	-0.120	0.904205	
healthFair	1051.3	789.0	1.332	0.182751	
healthGood	2766.0	777.4	3.558	0.000375	***
healthVery Good	4931.8	787.8	6.260	3.95e-10	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15530 on 16821 degrees of freedom

Multiple R-squared: 0.01646, Adjusted R-squared: 0.01616

F-statistic: 56.29 on 5 and 16821 DF, p-value: < 2.2e-16

Only 1.6% has been explained, so very poor!

Recap

- What are qualitative/categorical variables?
- What are dummy variables?
- Why are qualitative/categorical variables used?
- How to use qualitative/categorical variables in regression model?

Formative assessment support

≡ [ENVS225-202526](#) > Quizzes

2025-26 Academic Year

[Home](#)

[Announcements](#)

[Assignments](#)

[Grades](#)

[Modules](#)

[Discussions](#)

[Panopto Recordings](#)

[People](#)

[Reading Lists @
Liverpool](#)

[Search](#)

[Student Course
Feedback](#)

[Quizzes](#)

▼ Assignment quizzes



Quantitative Block: Topic, Research Question, Dataset.

Due 22 Oct at 23:59 1 pt



Annual Population Survey: Data Access Test

Due 2 Nov at 23:59 1 pt

▼ Practice quizzes



Week 1 - Quiz

15 pts 15 Questions



Week 2 - Quiz

9 pts 9 Questions

Formative assessment support

Quantitative Block: Topic, Research Questions, Dataset

1 Essay 1 point



Provide:

- Research topic/aim (1 sentence)
- 1 Research question
- Dataset(s)

Edit View Insert Format Tools Table

12pt ▾ Paragraph ▾ | **B** *I* U A ▾ ▾ T^2 ▾ | ▾ ▾ ▾ ▾ | ▾ ▾ ▾ | ▾ \sqrt{x}

p

|

Submit