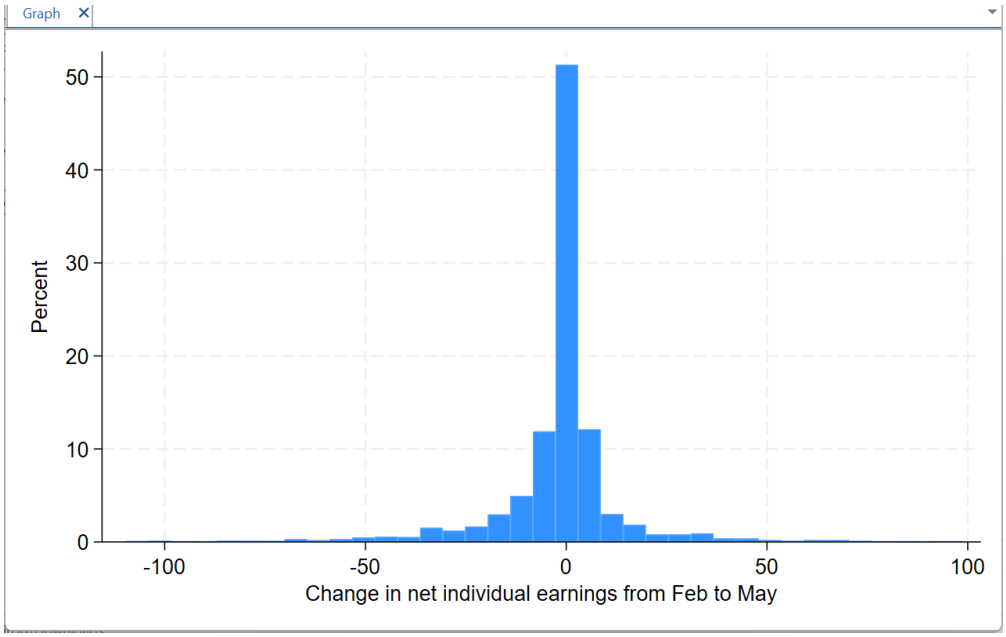
Student Number:100428943

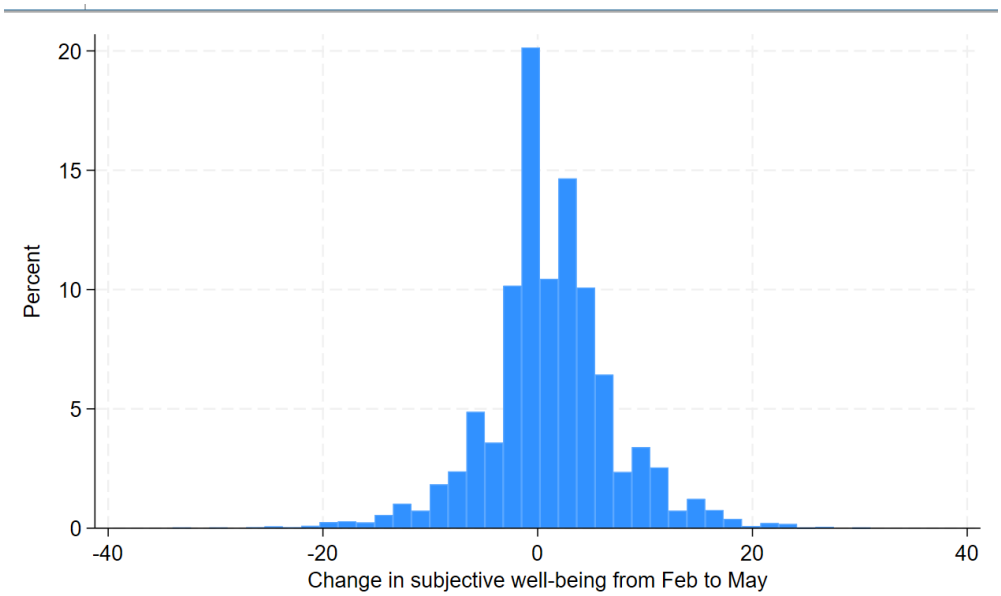
Data Analysis assignment

All stata code is in *red*

question A

```
generate dpay = netpay_may - netpay_feb
label variable dpay "Change in Net Pay (May - Feb)"
generate dghq = ghq_may - ghq_feb
label variable dghq "Change in Well-being (May - Feb)"
histogram dpay if dpay > -110 & dpay < 100, percent
histogram dghq if dghq > -110 & dghq < 100, percent
```





Student Number:100428943

```
summarize dpay, detail
summarize dghq, detail
```

Variable	Sample Size	Mean	Median	Std. Dev.	Min	25th %ile	IQR	75th %ile	Max
dpay	6,370	-2.09	0	39.91	-522.15	-3.29	5.09	1.81	563.00
dghq	7,893	1.13	1	6.02	-34.00	-2.00	6.00	4.00	31.00

The histogram for dpay reveals a skewed distribution with the majority of observations clustered around zero, reflecting minimal changes for most individuals. The average change in pay is -2.09, whereas the median is 0, suggesting that a significant number maintained their earnings, while a smaller group faced decreases. The long left tail (minimum: -522.15) highlights significant losses for some, likely due to job losses or reduced hours. The right tail (maximum: 563.00) indicates significant pay increases for a select few, possibly associated with critical or high-demand roles. The observed variability, indicated by a standard deviation of 39.91, is significant; however, the interquartile range of 5.09 suggests that the majority of changes were relatively small.

The histogram for dghq exhibits a distribution that is relatively symmetric, with a minor right skew observed. The average (1.13) and midpoint (1) indicate a modest rise in stress levels throughout the lockdown period. Outliers consist of a low of -34 (enhanced well-being) and a high of 31 (notable stress escalation). The results indicate that although there was a slight overall decline in well-being, individual experiences exhibited considerable variability, as evidenced by the standard deviation (6.02) and interquartile range (6).

When analysing the two variables, dpay shows a higher degree of variability and a more significant presence of outliers. The financial consequences of the lockdown were pronounced for certain individuals; however, the overall modest enhancement in well-being indicates that non-economic elements, like decreased commuting or increased family time, might have alleviated stress for a considerable number of people.

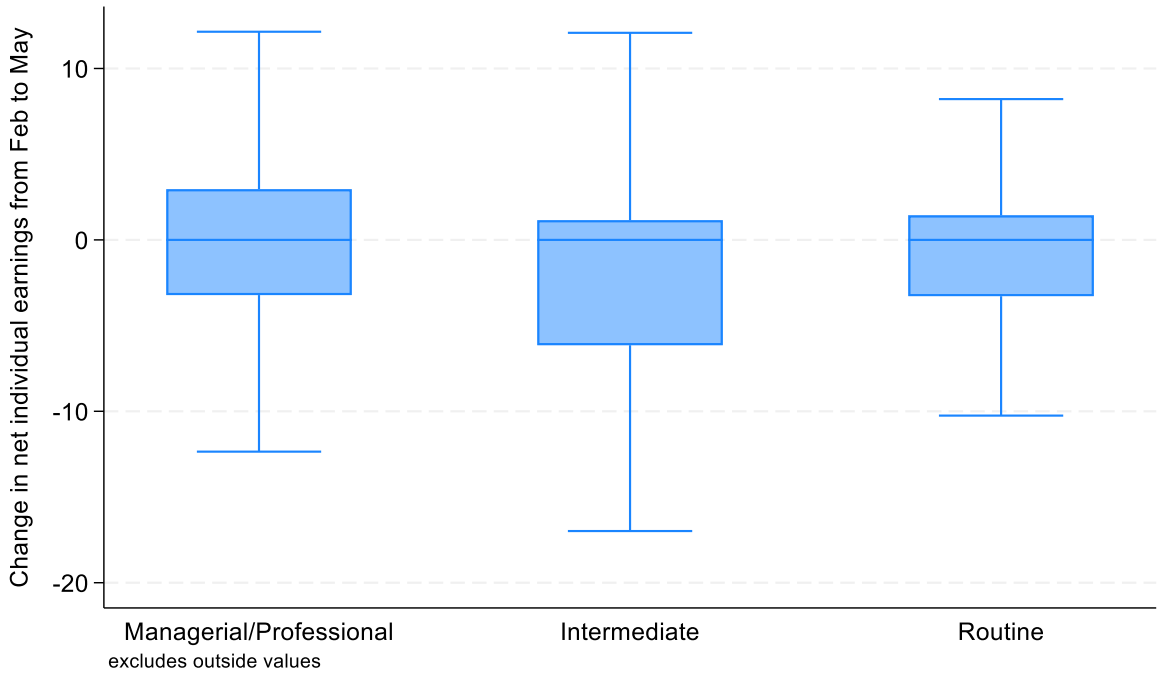
[252 words}

Question B

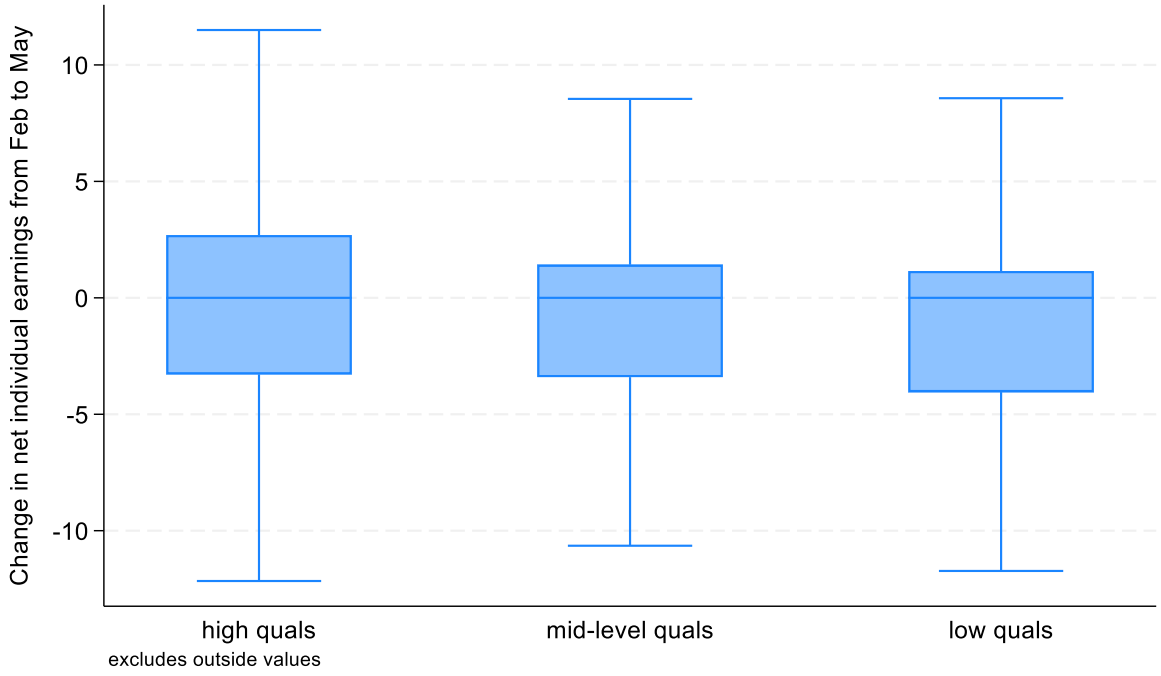
```
recode qual (1 2 = 1) (3 = 2) (4 5 9 = 3), gen(qual1)
label define qual1_labels 1 "high quals" 2 "mid-level quals" 3 "low quals"
label values qual1 qual1_labels
codebook qual1
```

```
qual1                     RECODE of qual (Highest Educational Qualification Attained, pre-Covid
-------------------------------------------------------------------------------------------------

                  Type: Numeric (byte)
                 Label: qual1

                 Range: [1,3]                         Units: 1
         Unique values: 3                         Missing .: 0/7,893

            Tabulation: Freq.   Numeric  Label
                        4,491         1  high quals
                        1,611         2  mid-level quals
                        1,791         3  low quals
```

```
graph box dpay, over(job_status) nooutsides
```




```
graph box dpay, over(qual1) nooutsides
```



The box plot for dpay by job status demonstrates variations in pay distribution during the lockdown. Every category shows a median of 0, indicating that the majority of individuals within each group saw no net change in their earnings. Nonetheless, the size of the boxes, represented by the interquartile ranges, exhibit variation, indicating disparities in compensation variability for the central 50%. Managerial and Intermediate roles show a

wider range, characterised by larger interquartile ranges and whiskers that extend further, signifying more pronounced fluctuations in wages. Routine jobs have the narrowest IQR and range, suggesting more uniform and stable pay changes. The lack of outliers suggests minimal significant variations. In general, positions of higher status encountered increased fluctuations, whereas routine occupations saw more stable pay outcomes.

The box plots for dpay categorised by qualification levels reveal consistent patterns among the groups, as each category exhibits a median of 0, suggesting that there is no net change in earnings for the majority of individuals.Individuals possessing the highest qualifications demonstrate significant variability, characterised by the largest interquartile range (IQR) and the broadest span between maximum and minimum values, indicating a wide array of fluctuations in pay throughout the lockdown period. Mid-level qualifications exhibit the smallest interquartile range, indicating the most consistent earnings, although their range implies some degree of variability. Individuals with lower qualifications are positioned between the two groups, exhibiting a moderate interquartile range and overall range. The absence of outliers indicates that significant pay variations were rare among all groups.

[249 words]

Question C

```
label define FEMALES 0 "male" 1 "female"
label values female FEMALES

label define URBAN "0" rural area "1" urban area
label values urban URBAN

ttest dghq, by(female)
```

```
Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
male	3,280	.8960366	.0970501	5.558186	.7057516	1.086322
female	4,613	1.301539	.0930295	6.318477	1.119157	1.483921
Combined	7,893	1.133029	.0677285	6.017169	1.000264	1.265795
diff		-.4055025	.1373639		-.6747721	-.136233

```
    diff = mean(male) - mean(female)                              t =  -2.9520
H0: diff = 0                                     Degrees of freedom =     7891

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0016         Pr(|T| > |t|) = 0.0032          Pr(T > t) = 0.9984
```

```
ttest dghq, by(urban)
```

```
Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
rural ar	1,938	1.111455	.1412836	6.219689	.8343712	1.388539
urban ar	5,955	1.14005	.0771076	5.950292	.9888914	1.291209
Combined	7,893	1.133029	.0677285	6.017169	1.000264	1.265795
diff		-.0285953	.1573701		-.3370822	.2798917

```
    diff = mean(rural ar) - mean(urban ar)                        t =  -0.1817
H0: diff = 0                                     Degrees of freedom =     7891

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.4279         Pr(|T| > |t|) = 0.8558          Pr(T > t) = 0.5721
```

.

Two-sample t-tests were employed to examine changes in well-being (dghq) based on gender (female) and location (urban). The mean well-being change for males was 0.896, whereas for females it was 1.302, resulting in a mean difference of -0.406. The p-value of 0.0032 demonstrates significance at the 5% level, confirming that females experienced a greater increase in stress. This variation underscores the distinct effects experienced by different genders during the lockdown, potentially associated with varying economic or social responsibilities.

In terms of location, rural residents experienced an average well-being change of 1.111, while urban residents had an average of 1.140, resulting in a mean difference of -0.029. The p-value of 0.8558 indicates that there is no significant difference observed. This indicates that geographic location had minimal impact on changes in well-being, as both rural and urban areas probably encountered comparable challenges, including limited mobility and economic uncertainty.

The findings suggest that gender played a significant role in the changes in well-being during the lockdown, whereas location did not exhibit a notable effect.

[171 words]

Question D

```
reg dpay female age ib3.qual  ib1.job_status
```

Student Number:100428943

```
. reg dpay female age ib3.qual1 ib1.job_status
```

Source	SS	df	MS		
Model	21629.3612	6	3604.89354	Number of obs	= 6,370
Residual	10124416.2	6,363	1591.13881	F(6, 6363)	= 2.27
Total	10146045.6	6,369	1593.03589	Prob > F	= 0.0347
				R-squared	= 0.0021
				Adj R-squared	= 0.0012
				Root MSE	= 39.889

dpay	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female	2.230024	1.020236	2.19	0.029	.2300176	4.230031
age	-.0435667	.0421903	-1.03	0.302	-.1262738	.0391405
qual1						
high quals	-2.203088	1.393524	-1.58	0.114	-4.934866	.528689
mid-level quals	-.665459	1.584069	-0.42	0.674	-3.770768	2.43985
job_status						
Intermediate	-2.339945	1.307259	-1.79	0.074	-4.902612	.2227224
Routine	.5957302	1.3492	0.44	0.659	-2.049156	3.240617
_cons	.3899336	2.609601	0.15	0.881	-4.725762	5.50563

The calculated coefficient for the female variable is 2.23, indicating that, on average, females see net earnings increase of £2.23 per day compared to males, assuming all other variables remain unchanged. To determine if this coefficient is significantly different from 0, we examine the p-value, which is 0.029. Given that the p-value is below the 0.05 significance threshold, we dismiss the null hypothesis (H_0: the coefficient for female equals 0). This suggests that the coefficient for female is notably distinct from 0, indicating a meaningful disparity in net earnings between females and males.

testparm i.job_status

```
 ( 1)  2.job_status = 0
 ( 2)  3.job_status = 0

       F(  2,  6363) =    2.34
            Prob > F =    0.0962
```

The F test for the job status dummies yields a F statistic of 2.34 and a p value of 0.0962. Since the p value exceeds 0.05, we are unable to reject the null hypothesis, indicating that the dummy variables do not have a joint significance in affecting dpay.

Reg dpay female age ib3.qual1 ib1.job_status, level (99)

```
. reg dpay female age ib3.qual1 ib1.job_status, level(99)
```

Source	SS	df	MS		
Model	21629.3612	6	3604.89354	Number of obs	= 6,370
Residual	10124416.2	6,363	1591.13881	F(6, 6363)	= 2.27
Total	10146045.6	6,369	1593.03589	Prob > F	= 0.0347
				R-squared	= 0.0021
				Adj R-squared	= 0.0012
				Root MSE	= 39.889

dpay	Coefficient	Std. err.	t	P>|t|	[99% conf.	interval]
female	2.230024	1.020236	2.19	0.029	-.3987185	4.858767
age	-.0435667	.0421903	-1.03	0.302	-.1522742	.0651409
qual1						
high quals	-2.203088	1.393524	-1.58	0.114	-5.793647	1.38747
mid-level quals	-.665459	1.584069	-0.42	0.674	-4.746974	3.416057
job_status						
Intermediate	-2.339945	1.307259	-1.79	0.074	-5.708231	1.028341
Routine	.5957302	1.3492	0.44	0.659	-2.880622	4.072082
_cons	.3899336	2.609601	0.15	0.881	-6.333969	7.113836

The latest regression analysis indicates that the 99% confidence interval for the age coefficient implies that the impact of ageing by one year on daily net earnings, from February to May 2020, may vary from a reduction of £0.15 to an increase of £0.07, assuming all other factors remain unchanged. The p-value for age is 0.302, which exceeds the 0.05 significance threshold, reinforcing the conclusion that the impact of age on daily pay is not significant. Given that this interval encompasses zero, we are unable to assert with confidence that age significantly influences daily pay (dpay). The outcome may manifest as positive, negative, or negligible. The overall model presents a low R-squared value of 0.0021, suggesting that a minimal amount of the variation in daily pay is accounted for by the variables included. This implies that there may be additional unobserved factors, such as experience, location, or industry, that could be affecting daily earnings. Consequently, we determine that age does not exert a significant influence on net earnings within this model.

[313 words]

Question E

```
gen lpay_feb = log(netpay_feb)
reg dpay female age ib3.qual1 ib1 job_status lpay_feb
```

```
. reg dpay female age ib3.qual1 ib1.job_status lpay_feb
```

Source	SS	df	MS		
Model	583038.18	7	83291.1686	Number of obs	= 6,349
Residual	9431892.29	6,341	1487.44556	F(7, 6341)	= 56.00
				Prob > F	= 0.0000
				R-squared	= 0.0582
				Adj R-squared	= 0.0572
Total	10014930.5	6,348	1577.6513	Root MSE	= 38.567

dpay	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female	-3.600854	1.034332	-3.48	0.001	-5.628495	-1.573212
age	-.0744539	.0409371	-1.82	0.069	-.1547045	.0057968
qual1						
high quals	1.586854	1.362355	1.16	0.244	-1.083822	4.257529
mid-level quals	.5711636	1.535271	0.37	0.710	-2.438486	3.580813
job_status						
Intermediate	-8.76671	1.309586	-6.69	0.000	-11.33394	-6.199479
Routine	-6.983139	1.36559	-5.11	0.000	-9.660158	-4.306119
lpay_feb	-15.25448	.7851199	-19.43	0.000	-16.79358	-13.71538
_cons	66.11573	4.25313	15.55	0.000	57.77816	74.4533

The coefficient for lpay_feb is -15.25448, signifying that a 1% rise in netpay_feb correlates with a 15.25 unit reduction in dpay, assuming all other variables remain constant. The p-value is 0.000, indicating statistical significance.

The estimated coefficient for female was 2.23; however, it turned negative, decreasing to -3.60 following the inclusion of lpay_feb. A comparable change was observed for the job status dummies: Intermediate fell from -2.34 to -8.77, while Routine shifted from 0.60 to -6.98. Incorporating lpay_feb allows us to factor in earnings immediately preceding the lockdown, which had not been taken into account before. This modification enables a more precise evaluation of the supplementary effects of gender and specific job roles on variations in daily earnings from February to May. However, it indicates that throughout the lockdown period, identifying as female or holding a routine or intermediate job role correlated with a reduction in net daily earnings.

Significant alterations in the coefficients were expected since netpay_feb is probably linked to both gender and job status. By incorporating lpay_feb, we tackle the possible omitted variable bias that initially assigned some of the impact to female or job-status factors instead of baseline earnings. In doing so, we obtain a clearer understanding of how these factors independently influence dpay once prior pay levels are controlled for. This highlights the significance of considering pre-lockdown wages when evaluating the varying effects of the pandemic on different demographic and occupational categories.

[237 words]

Question F

Ln (150) = 5.0106352941

```
margins, at (age=30 female=1 job_status=1 qual1=1 lpay_feb=5.0106)
```

```
Adjusted predictions                                  Number of obs = 6,349
Model VCE: OLS

Expression: Linear prediction, predict()
At: female     =        1
    age        =       30
    qual1      =        1
    job_status =        1
    lpay_feb   = 5.010635
```

	Margin	Delta-method std. err.	t	P>|t|	[95% conf.	interval]
_cons	-14.5665	1.214907	-11.99	0.000	-16.94813	-12.18487

Ln (85) = 4.44265125649

```
margins, at (age45 female=0 job_status=3 qual1=3 lpay_feb=4.44265125649)
```

```
. margins, at(age=45 female=0 job_status=3 qual1=3 lpay_feb=4.44265125649)

Adjusted predictions                                      Number of obs = 6,349
Model VCE: OLS

Expression: Linear prediction, predict()
At: female     =        0
    age        =       45
    qual1      =        3
    job_status =        3
    lpay_feb   = 4.442651
```

	Delta-method Margin	std. err.	t	P>|t|	[95% conf.	interval]
_cons	-11.98815	1.447796	-8.28	0.000	-14.82632	-9.149978

Ln (105) = 4.65396035016

margins, at (age=60 female=1 job_status=2 qual1=2 lpay_feb=4.65396035016)

```
. margins, at(age=60 female=1 job_status=2 qual1=2 lpay_feb=4.65396035016)

Adjusted predictions                                      Number of obs = 6,349
Model VCE: OLS

Expression: Linear prediction, predict()
At: female     =        1
    age        =       60
    qual1      =        2
    job_status =        2
    lpay_feb   = 4.65396
```

	Delta-method Margin	std. err.	t	P>|t|	[95% conf.	interval]
_cons	-21.14162	1.828323	-11.56	0.000	-24.72575	-17.55749

In Case 1, the predicted dpay for a 30-year-old female with a managerial/professional job status, high qualifications, and a net pay of £150 per day in February 2020 is a decrease of £14.57.

In Case 2, the predicted dpay for a 45-year-old male with a routine job status, low qualifications, and a net pay of £85 per day in February 2020 is a decrease of £11.99.

In Case 3, the predicted dpay for a 60-year-old female with an intermediate job status, mid-level qualifications, and a net pay of £105 per day in February 2020 is a decrease of £21.14.

[99 words]

Question G

```
generate age2 = age^2
reg dghq female age ib3.qual1 ib3.health ib6.region
```

Student Number:100428943

Source	SS	df	MS	Number of obs	=	7,893
				F(19, 7873)	=	5.97
Model	4056.19999	19	213.48421	Prob > F	=	0.0000
Residual	281684.119	7,873	35.7784986	R-squared	=	0.0142
				Adj R-squared	=	0.0118
Total	285740.319	7,892	36.2063253	Root MSE	=	5.9815

dghq	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female	.3843285	.1375386	2.79	0.005	.1147164	.6539406
age	-.0122157	.0054321	-2.25	0.025	-.022864	-.0015673
qual1						
high quals	.0681938	.1708361	0.40	0.690	-.2666904	.4030779
mid-level quals	.0807164	.2081538	0.39	0.698	-.3273203	.4887531
health						
Excellent	.6802375	.2271958	2.99	0.003	.2348734	1.125602
Very good	.4109536	.1576186	2.61	0.009	.1019792	.7199279
Fair	-.8116821	.2327405	-3.49	0.000	-1.267915	-.3554489
Poor	-3.347395	.5422311	-6.17	0.000	-4.410312	-2.284478
region						
North East	.9023327	.4248388	2.12	0.034	.0695359	1.73513
North West	.5275506	.3029104	1.74	0.082	-.066234	1.121335
Yorkshire and the Humber	.6028728	.3171311	1.90	0.057	-.0187882	1.224534
East Midlands	.1125564	.324182	0.35	0.728	-.5229264	.7480391
West Midlands	.4781575	.3145893	1.52	0.129	-.138521	1.094836
London	.6736006	.3037581	2.22	0.027	.0781541	1.269047
South East	.3233691	.2775178	1.17	0.244	-.2206394	.8673777
South West	.141969	.3038655	0.47	0.640	-.4536881	.737626
Wales	.4982558	.3524244	1.41	0.157	-.1925895	1.189101
Scotland	.2937282	.3110396	0.94	0.345	-.315992	.9034485
Northern Ireland	.0828161	.3942853	0.21	0.834	-.6900878	.8557199
_cons	.9422807	.3909363	2.41	0.016	.1759417	1.70862

reg dghq female ib3.qual1 ib3.health ib6.region c.age##c.age

Source	SS	df	MS	Number of obs	=	7,893
				F(20, 7872)	=	5.82
Model	4161.94793	20	208.097397	Prob > F	=	0.0000
Residual	281578.371	7,872	35.7696102	R-squared	=	0.0146
				Adj R-squared	=	0.0121
Total	285740.319	7,892	36.2063253	Root MSE	=	5.9808

dghq	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female	.3796145	.1375488	2.76	0.006	.1099823	.6492466
qual1						
high quals	.0938486	.1714653	0.55	0.584	-.242269	.4299661
mid-level quals	.0715883	.2081956	0.34	0.731	-.3365304	.479707
health						
Excellent	.6806433	.2271677	3.00	0.003	.2353342	1.125952
Very good	.4068436	.1576172	2.58	0.010	.0978721	.7158151
Fair	-.8087312	.2327179	-3.48	0.001	-1.26492	-.3525422
Poor	-3.329763	.5422608	-6.14	0.000	-4.392738	-2.266788
region						
North East	.9116013	.4248203	2.15	0.032	.0788409	1.744362
North West	.5297625	.3028755	1.75	0.080	-.0639538	1.123479
Yorkshire and the Humber	.6079553	.3171054	1.92	0.055	-.0136555	1.229566
East Midlands	.1163164	.3241491	0.36	0.720	-.5191019	.7517346
West Midlands	.4762673	.3145521	1.51	0.130	-.1403383	1.092873
London	.6718632	.3037221	2.21	0.027	.0764874	1.267239
South East	.3141533	.2775351	1.13	0.258	-.2298892	.8581958
South West	.1355875	.3038504	0.45	0.655	-.46004	.731215
Wales	.4988519	.3523808	1.42	0.157	-.191908	1.189612
Scotland	.296599	.3110055	0.95	0.340	-.3130543	.9062523
Northern Ireland	.0878068	.394247	0.22	0.824	-.685022	.8606356
age	-.0664913	.0320304	-2.08	0.038	-.1292794	-.0037033
c.age#c.age	.0006095	.0003545	1.72	0.086	-.0000854	.0013045
_cons	2.042143	.7496517	2.72	0.006	.5726269	3.51166

The linear representation of age is favoured in this regression model due to insufficient evidence indicating a nonlinear association between age and variations in subjective well-being. In the linear model, the coefficient for age is significant (−0.0122, p = 0.025), indicating a clear negative relationship: as age increases, dghq decreases. The R-squared value for this model is 0.0118, indicating the extent to which the predictors account for the variance. Although the explanatory power is limited, the straightforward nature of the linear relationship facilitates interpretation, particularly in light of the absence of evidence supporting a more intricate relationship.

In the quadratic model, both the age term (-0.0664, p = 0.038) and the squared term

(0.000695, p = 0.006) show significance; however, the enhancement in model fit is negligible, as indicated by a slight increase in R-squared to 0.0146. This minor adjustment indicates that incorporating the squared term does not significantly improve the model's capacity to account for variations in dghq.

The other predictors in the model show a high degree of consistency across both the linear and quadratic specifications. For example, the "female" variable shows significance in both models (coefficients of 0.3843 and 0.3796, $p < 0.01$), and the health and regional variables display comparable patterns. This consistency reinforces the notion that the additional complexity of the quadratic form for age is unwarranted. Consequently, the linear representation of age, which reflects the overall negative trend without excessive fitting, is the more suitable option.

[242 words]

Question H

```
test 1.region 2.region 3.region 4.region 5.region 7.region 8.region 9.region
10.region 11.region 12.region
```

```
( 1)  1.region = 0
( 2)  2.region = 0
( 3)  3.region = 0
( 4)  4.region = 0
( 5)  5.region = 0
( 6)  7.region = 0
( 7)  8.region = 0
( 8)  9.region = 0
( 9)  10.region = 0
(10)  11.region = 0
(11)  12.region = 0

      F( 11,  7872) =    1.12
           Prob > F =    0.3423

.
```

The coefficient for “North East” (0.9023, p = 0.032) indicates that individuals in this region have a dghq score 0.9023 units higher than those in the East of England, and this finding is significant. The coefficient for “London” (0.6738, p = 0.054) suggests a 0.6738-unit increase in the dghq score, though it is only marginally significant at the 10% level.

The joint F-test for all regional dummies (F = 2.85, p = 0.0015) indicates that the regional variables, when considered together, exert a statistically significant influence on’ dghq’. This differs from individual assessments, as the collective evaluation considers the overall impact of all areas. While individual regions such as “London” may not exhibit strong significance.

The separate tests for "North East" and "London" exhibit different outcomes in comparison to the combined test. The coefficient for "North East" (0.9023, p = 0.032) is statistically significant, suggesting that persons in this location report a 0.9023-unit higher dghq score compared to those in the East of England. The coefficient for "London" (0.6738, p = 0.054) is marginally significant, indicating a 0.6738-unit increase in the dghq score.

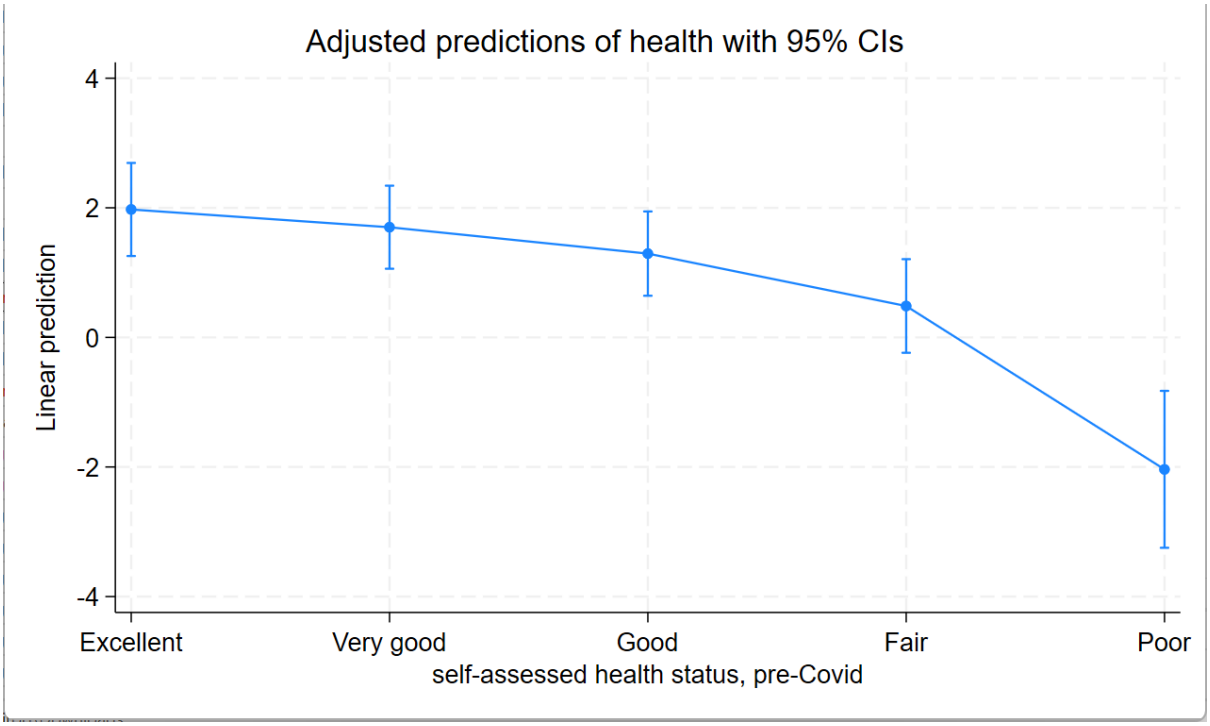
The combined F-test for all regional dummies (F = 1.12, p = 0.3423) does not reject the null hypothesis, indicating no significant collective effect of regions on dghq. This difference indicates that whereas some regions exhibit individual significance, their overall impact is diminished, probably because to disparities among all regions. This underscores the significance of evaluating both individual and collective effects.

[244 words]

Question I

```
margins health, at(age=24 female=1 qual1=3 region=4)
marginsplot
```

Student Number:100428943




```
Adjusted predictions                                  Number of obs = 7,893
Model VCE: OLS

Expression: Linear prediction, predict()
At: female =  1
    qual1  =  3
    region =  4
    age    = 24
```

I picked Age 24 as it captures a young adult group where health is less influenced by age-related health declines. Female gender was selected to focus on a significant subgroup, acknowledging that well-being patterns can vary by gender. High qualifications were chosen as they often correlate with better social economic outcomes, helping to isolate the direct impact of health on well-being by minimising confounding effects related to education. The East Midlands region was selected as a mid-level region that avoids extremes in urbanisation, income, or regional disparities.

The model's results were utilised to calculate the expected $dghq$ values for each health status category. The data indicates a distinct decline in dghq as health declines. Individuals in "Excellent" health have the highest anticipated dghq score, roughly 2 units, whilst those in "Very Good" and "Good" health demonstrate marginally lower values of about 1.5 and 1 unit, respectively. The dghq score persistently decreases for those with "Fair" health (approaching 0) and experiences a significant reduction for those with "Poor" health, with a predicted value of roughly -3. This underscores a notable correlation between deteriorating health status and decreasing dghq, with the most pronounced fall evident between "Fair" and "Poor" health. The results are consistent across fixed demographic and regional variables, offering vital insight into the impact of health status on dghq

Student Number:100428943

[219 words]

Question J

question e model

```
reg dpay female age ib3.qual1 ib1.job_status lpay_feb
estat hettest, rhs fstat
```

Source	SS	df	MS	Number of obs	=	6,349
				F(7, 6341)	=	56.00
Model	583038.18	7	83291.1686	Prob > F	=	0.0000
Residual	9431892.29	6,341	1487.44556	R-squared	=	0.0582
				Adj R-squared	=	0.0572
Total	10014930.5	6,348	1577.6513	Root MSE	=	38.567

dpay	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female	-3.600854	1.034332	-3.48	0.001	-5.628495	-1.573212
age	-.0744539	.0409371	-1.82	0.069	-.1547045	.0057968
qual1						
high quals	1.586854	1.362355	1.16	0.244	-1.083822	4.257529
mid-level quals	.5711636	1.535271	0.37	0.710	-2.438486	3.580813
job_status						
Intermediate	-8.76671	1.309586	-6.69	0.000	-11.33394	-6.199479
Routine	-6.983139	1.36559	-5.11	0.000	-9.660158	-4.306119
lpay_feb	-15.25448	.7851199	-19.43	0.000	-16.79358	-13.71538
_cons	66.11573	4.25313	15.55	0.000	57.77816	74.4533

```
Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: All independent variables

H0: Constant variance

F(7, 6341) =  14.89
  Prob > F = 0.0000
```

question g model

```
reg dghq female age ib3.qual1 ib3.health ib6.region
estat hettest, rhs fstat
```

Student Number:100428943

dghq	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
female	.3843285	.1375386	2.79	0.005	.1147164	.6539406
age	-.0122157	.0054321	-2.25	0.025	-.022864	-.0015673
qual1						
high quals	.0681938	.1708361	0.40	0.690	-.2666904	.4030779
mid-level quals	.0807164	.2081538	0.39	0.698	-.3273203	.4887531
health						
Excellent	.6802375	.2271958	2.99	0.003	.2348734	1.125602
Very good	.4109536	.1576186	2.61	0.009	.1019792	.7199279
Fair	-.8116821	.2327405	-3.49	0.000	-1.267915	-.3554489
Poor	-3.347395	.5422311	-6.17	0.000	-4.410312	-2.284478
region						
North East	.9023327	.4248388	2.12	0.034	.0695359	1.73513
North West	.5275506	.3029104	1.74	0.082	-.066234	1.121335
Yorkshire and the Humber	.6028728	.3171311	1.90	0.057	-.0187882	1.224534
East Midlands	.1125564	.324182	0.35	0.728	-.5229264	.7480391
West Midlands	.4781575	.3145893	1.52	0.129	-.138521	1.094836
London	.6736006	.3037581	2.22	0.027	.0781541	1.269047
South East	.3233691	.2775178	1.17	0.244	-.2206394	.8673777
South West	.141969	.3038655	0.47	0.640	-.4536881	.737626
Wales	.4982558	.3524244	1.41	0.157	-.1925895	1.189101
Scotland	.2937282	.3110396	0.94	0.345	-.315992	.9034485
Northern Ireland	.0828161	.3942853	0.21	0.834	-.6900878	.8557199
_cons	.9422807	.3909363	2.41	0.016	.1759417	1.70862

```
.
. estat hettest, rhs fstat

Breusch-Pagan/Cook-Weisberg test for heteroskedasticity
Assumption: i.i.d. error terms
Variables: All independent variables

H0: Constant variance

F(19, 7873) =  12.56
   Prob > F = 0.0000
```

question e reestimate

```
reg dpay female age ib3.qual1 ib1.job_status lpay_feb, vce(robust)
```

```
Linear regression                               Number of obs     =      6,349
                                                F(7, 6341)        =      12.11
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0582
                                                Root MSE          =     38.567
```

dpay	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
female	-3.600854	1.273074	-2.83	0.005	-6.096509	-1.105198
age	-.0744539	.0408017	-1.82	0.068	-.154439	.0055313
qual1						
high quals	1.586854	1.246769	1.27	0.203	-.8572346	4.030942
mid-level quals	.5711636	1.507391	0.38	0.705	-2.383832	3.526159
job_status						
Intermediate	-8.76671	1.33966	-6.54	0.000	-11.3929	-6.140524
Routine	-6.983139	1.588709	-4.40	0.000	-10.09755	-3.868731
lpay_feb	-15.25448	1.927765	-7.91	0.000	-19.03355	-11.4754
_cons	66.11573	8.594212	7.69	0.000	49.26817	82.96329

question g reesimiate

```
reg dghq female age ib3.qual1 ib3.health ib6.region, vce(robust)
```

dghq	Coefficient	std. err.	t	P>|t|	[95% conf.	interval]
female	.3843285	.1348445	2.85	0.004	.1199976	.6486595
age	-.0122157	.0053829	-2.27	0.023	-.0227676	-.0016638
qual1						
high quals	.0681938	.1690473	0.40	0.687	-.2631838	.3995713
mid-level quals	.0807164	.2026116	0.40	0.690	-.3164562	.477889
health						
Excellent	.6802375	.2158036	3.15	0.002	.2572052	1.10327
Very good	.4109536	.1535751	2.68	0.007	.1099057	.7120014
Fair	-.8116821	.2782023	-2.92	0.004	-1.357032	-.2663318
Poor	-3.347395	.7757526	-4.32	0.000	-4.868076	-1.826714
region						
North East	.9023327	.4214918	2.14	0.032	.076097	1.728568
North West	.5275506	.2964685	1.78	0.075	-.0536062	1.108708
Yorkshire and the Humber	.6028728	.3104835	1.94	0.052	-.0057573	1.211503
East Midlands	.1125564	.3130433	0.36	0.719	-.5010916	.7262043
West Midlands	.4781575	.3225586	1.48	0.138	-.154143	1.110458
London	.6736006	.3054262	2.21	0.027	.0748841	1.272317
South East	.3233691	.2725263	1.19	0.235	-.2108546	.8575929
South West	.141969	.300353	0.47	0.636	-.4468025	.7307405
Wales	.4982558	.3563125	1.40	0.162	-.2002112	1.196723
Scotland	.2937282	.3289848	0.89	0.372	-.3511692	.9386257
Northern Ireland	.0828161	.379519	0.22	0.827	-.6611419	.826774
_cons	.9422807	.390419	2.41	0.016	.1769559	1.707605

Employing the Breusch-Pagan test, I obtained an F-statistic of 14.49 and a p-value of 0.0000 for the model in Question E. The results indicate that the variance of the residuals is not constant across all observations, allowing us to reject the null hypothesis of homoskedasticity. We arrive at the identical conclusion for model G, which produced an F-statistic of 12.56 and a p-value of 0.0000. Consequently, the tests clearly indicate the presence of heteroskedasticity in both models. To address this issue, both models were re-estimated using heteroskedasticity-robust standard errors (RSEs). In model E, while the coefficients remained unchanged, adjustments in standard errors slightly impacted the significance of variables, such as “female,” where the standard error increased from 1.0343 to 1.2731, but the variable remained significant (p = 0.005). Similarly, in model G, the standard error for “age” increased from 0.0054 to 0.0059, yet it remained significant (p = 0.023). Overall, robust standard errors ensured the reliability of the results despite the presence of heteroskedasticity.

[164 words]

Question K

```
reg dghq female age ib3.qual1 ib3.health ib6.region i.female#i.qual1, vce(robust)
```

Linear regression

Number of obs	7,893
F(21, 7871)	4.26
Prob > F	0.0000
R-squared	0.0143
Root MSE	5.982

dghq	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
female	.1776794	.2682407	0.66	0.508	-.3481437	.7035024
age	-.0121767	.0053894	-2.26	0.024	-.0227413	-.0016121
qual1						
high quals	-.1006749	.2316006	-0.43	0.664	-.5546735	.3533237
mid-level quals	-.0339089	.2717953	-0.12	0.901	-.5666999	.4988821
health						
Excellent	.6786004	.2159589	3.14	0.002	.2552636	1.101937
Very good	.4093736	.1535934	2.67	0.008	.1082898	.7104573
Fair	-.8115769	.2782835	-2.92	0.004	-1.357086	-.2660674
Poor	-3.349017	.7760334	-4.32	0.000	-4.870249	-1.827786
region						
North East	.893984	.4218905	2.12	0.034	.0669666	1.721001
North West	.5243718	.2964566	1.77	0.077	-.0567619	1.105505
Yorkshire and the Humber	.6013037	.3106629	1.94	0.053	-.0076779	1.210285
East Midlands	.1080671	.3132249	0.35	0.730	-.5059368	.722071
West Midlands	.4773346	.322611	1.48	0.139	-.1550686	1.109738
London	.6751097	.3054086	2.21	0.027	.0764278	1.273792
South East	.3227791	.2726277	1.18	0.236	-.2116436	.8572018
South West	.1418805	.3004208	0.47	0.637	-.447024	.7307851
Wales	.4990613	.3562011	1.40	0.161	-.1991874	1.19731
Scotland	.2914069	.3290367	0.89	0.376	-.3535923	.9364061
Northern Ireland	.0838116	.379563	0.22	0.825	-.6602326	.8278558
female#qual1						
female#high quals	.2926205	.3245657	0.90	0.367	-.3436145	.9288555
female#mid-level quals	.2018743	.3921815	0.51	0.607	-.5669056	.9706541
female#low quals	0	(omitted)				
_cons	1.060103	.4052968	2.62	0.009	.2656138	1.854592

The interaction between female and qual1 was chosen to investigate whether the effect of gender on dghq varies by educational qualifications. This hypothesis is motivated by the idea that different gendered individuals may face different society biases with their different levels of qualifications. For instance, those with higher qualifications may have fewer disparities between male and female but they could be more present with those with lower qualifications. Females coefficient is 0.178 meaning that females have a 0.178 unit higher score than males. We can also see that the p value is greater than 0.05 at 0.508 thus there is no major direct association between gender and dghq when qualifications are at base level however, from the coefficient, you can see that females are slightly more stressed

Similar results are seen between the interaction terms, as females with high qualifications have a p value of 0.367 and females with mid-level qualifications have a p value of 0.607, both are statistically insignificant suggesting that being female stress levels do not differ across different qualification levels

```
testparm i.female#i.qual1
```

```
 ( 1)  1.female#1.qual1 = 0
 ( 2)  1.female#2.qual1 = 0

       F(  2,  7871) =    0.41
            Prob > F =    0.6660
```

My null hypothesis is the interaction terms for female and qual1 have no effect on dghq. My alternative hypothesis is that at least one of the interaction terms has a significant effect on dghq. The F-test gave us a very low f statistic against the null hypothesis with 0.6660 p value verifying its weakness as the p value is significantly larger than significant levels. Neither of the interaction terms significantly contribute to explaining variation in dghq thus we fail to reject the null hypotheses.

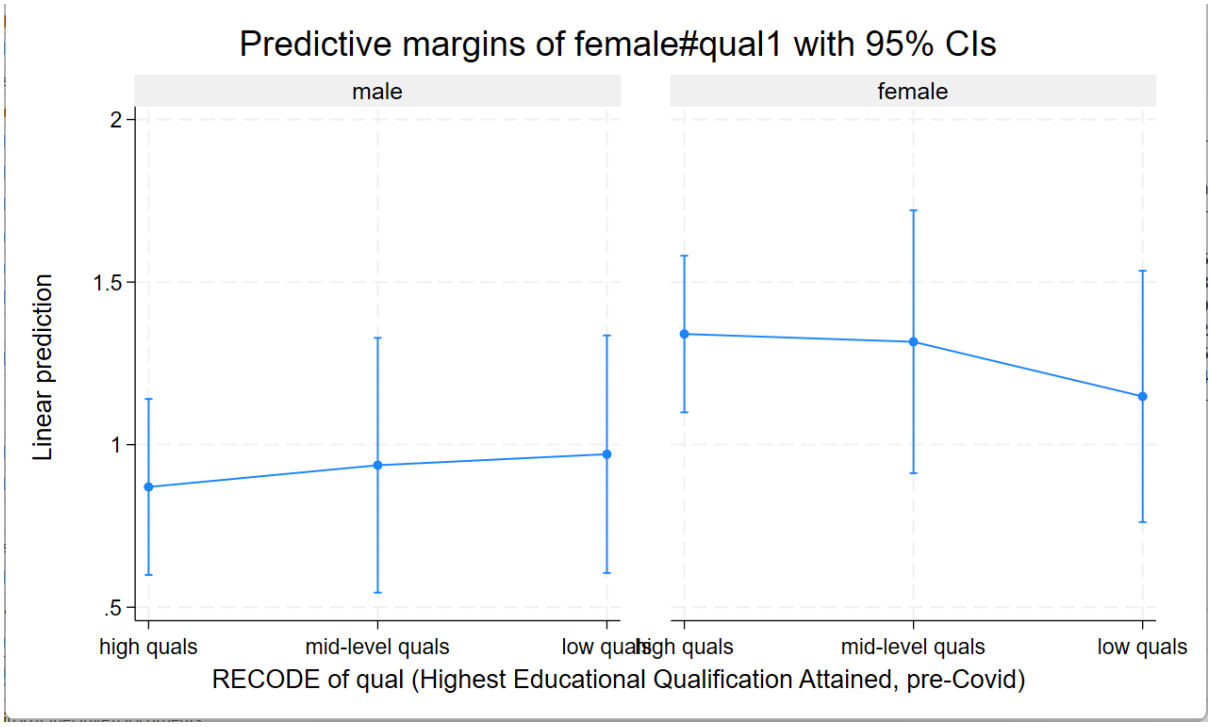
```
margins i.female#i.qual1
```

```
Predictive margins
Model VCE: Robust

Expression: Linear prediction, predict()
```

```
marginsplot, xdimension(qual1) by(female)
```

Student Number:100428943

	Margin	Delta-method std. err.	t	P>|t|	[95% conf.	interval]
female#qual1						
male#high quals	.8700092	.1380648	6.30	0.000	.5993655	1.140653
male#mid-level quals	.9367752	.199853	4.69	0.000	.5450102	1.32854
male#low quals	.9706841	.1863434	5.21	0.000	.6054015	1.335967
female#high quals	1.340309	.1228925	10.91	0.000	1.099407	1.581211
female#mid-level quals	1.316329	.2062573	6.38	0.000	.9120097	1.720648
female#low quals	1.148363	.1972318	5.82	0.000	.7617368	1.53499



I created a graph to demonstrate the different qualifications each gender has with their corresponding stress level. The graph shows that females across all three qualification levels experienced more stress during the lockdown than men. Between February to May, men with the lowest qualification type experienced the most stress in contrast to females with the highest qualification type reaching the highest level of stress however the gap between those are quite large, shown on the graph. Furthermore the CI's are relatively wide thus these predications are relatively imprecisely estimated

[347 words]

Student Number:100428943

Question L

```
reg furloughed_may c.age##c.age female ib8.region ib3.qual1 ib1.ethnic ib4.marstat
```

Source	SS	df	MS	Number of obs	=	7,221
				F(24, 7196)	=	9.40
Model	27.3940829	24	1.14142012	Prob > F	=	0.0000
Residual	873.468679	7,196	.121382529	R-squared	=	0.0304
				Adj R-squared	=	0.0272
Total	900.862761	7,220	.124773236	Root MSE	=	.3484

furloughed_may	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	-.0059772	.0021625	-2.76	0.006	-.0102163	-.0017381
c.age#c.age	.0000477	.0000233	2.04	0.041	1.92e-06	.0000934
female	-.0052261	.0084598	-0.62	0.537	-.0218099	.0113576
region						
North East	-.0266182	.0250842	-1.06	0.289	-.0757907	.0225543
North West	-.0059765	.0170224	-0.35	0.726	-.0393454	.0273925
Yorkshire and the Humber	.0372219	.0177831	2.09	0.036	.0023617	.072082
East Midlands	.0321167	.0183255	1.75	0.080	-.0038066	.0680401
West Midlands	.0246072	.0179522	1.37	0.171	-.0105843	.0597987
East of England	.0231602	.0167516	1.38	0.167	-.009678	.0559983
London	-.0010086	.0177692	-0.06	0.955	-.0358415	.0338244
South West	.0265026	.0168842	1.57	0.117	-.0065954	.0596006
Wales	-.0245926	.0202372	-1.22	0.224	-.0642635	.0150782
Scotland	.0013247	.0174544	0.08	0.940	-.0328911	.0355406
Northern Ireland	-.028124	.0235535	-1.19	0.232	-.0742958	.0180478
qual1						
high quals	-.1091237	.0103987	-10.49	0.000	-.1295081	-.0887392
mid-level quals	-.0377025	.0127051	-2.97	0.003	-.0626083	-.0127967
ethnic						
Mixed	-.0503195	.0317168	-1.59	0.113	-.1124937	.0118548
Asian	-.0606856	.0188499	-3.22	0.001	-.097637	-.0237342
Black	-.0531159	.0336774	-1.58	0.115	-.1191335	.0129017
Other	-.0728704	.0661355	-1.10	0.271	-.2025153	.0567745
marstat						
Married	-.0288916	.0156151	-1.85	0.064	-.0595018	.0017186
Living as couple	.0007402	.018812	0.04	0.969	-.0361369	.0376173
Widowed	-.027036	.0397055	-0.68	0.496	-.1048704	.0507984
Never married	-.0060828	.0195145	-0.31	0.755	-.0443369	.0321712
_cons	.4014645	.0536816	7.48	0.000	.2962328	.5066962

The coefficient for age is -0.0059772, with a p-value of 0.006, signifying a statistically significant inverse correlation between age and the probability of being furloughed in May. This indicates that for each extra year of age, the likelihood of being furloughed diminishes by roughly 0.6%, assuming all other variables remain unchanged. The coefficient for the squared term age^2 is 0.0000477, accompanied by a p-value of 0.041, suggesting a minor curvature in the connection. The positive coefficient of age^2 indicates that the adverse impact of age on furloughing decreases marginally with older individuals, resulting in a non-linear relationship. Generally, younger persons exhibit a higher propensity for furlough, although this tendency stabilises to some extent with advancing age.

Student Number:100428943

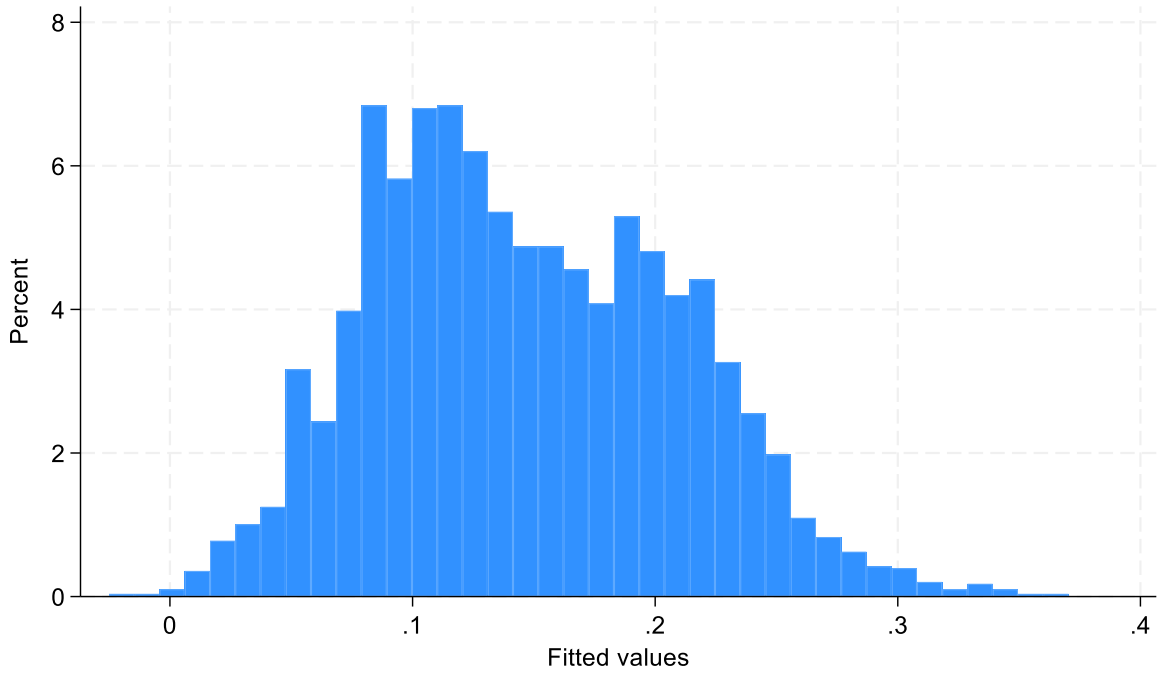
I included: ethnicity to evaluate potential disparities in furlough probabilities among ethnic groups resulting from variations in work sectors and structural imbalances;marital Status as it is relevant as family structure and financial stability may affect the probability of being furloughed, yielding distinct outcomes for single and married individuals; High qualifications as they are associated with job security and remote work prospects, hence diminishing the probability of furlough during economic downturns.

```
testparm i.marstat
```

```
. testparm i.marstat

 ( 1)  1.marstat = 0
 ( 2)  2.marstat = 0
 ( 3)  3.marstat = 0
 ( 4)  5.marstat = 0

       F(  4,  7196) =    2.25
            Prob > F =    0.0608
```

Marital status is classified as a categorical variable, with "Single" designated as the reference category. The coefficient for "Married" is -0.0289, signifying that married individuals are 2.89% less likely to be furloughed than their single counterparts, assuming other variables remain constant. Nevertheless, the p-value of 0.064 indicates that this result is only marginally significant at the 10% threshold. Alternative marital statuses, including "Living as a couple," "Widowed," and "Never married," exhibit analogous trends but do not possess individual statistical significance. The joint significance test for marital status (testparm i.marstat) produces an F-statistic of 2.25 and a p-value of 0.0608, indicating that marital status is marginally significant at the 10% level.

Student Number:100428943



The histogram showing the predicted probabilities from the Linear Probability Model (LPM) reveals that the majority of predicted values are between 0.1 and 0.3, and there are no values above 0.4. The model seems to generate probabilities that make sense, remaining within the expected limits of 0 and 1 for the dependent variable. But, having probabilities concentrated in a small range might mean there's not much variability in predictions. This could make it harder for the model to identify extreme outcomes or notable differences among observations. Even though the LPM seems to provide reliable probabilities here, it's important to be cautious when interpreting the results since LPMs can occasionally yield probabilities that fall outside the [0, 1] range in different datasets. The predictions seem to be in a reasonable range and show a decent level of reliability.

[433 words]

Student Number:100428943

Bibliography

University of Essex, Institute for Social and Economic Research. (2021). Understanding Society: COVID-19 Study, 2020-2021. [data collection]. 10th Edition. UK Data Service. SN: 8644, 10.5255/UKDA-SN-8644-10.