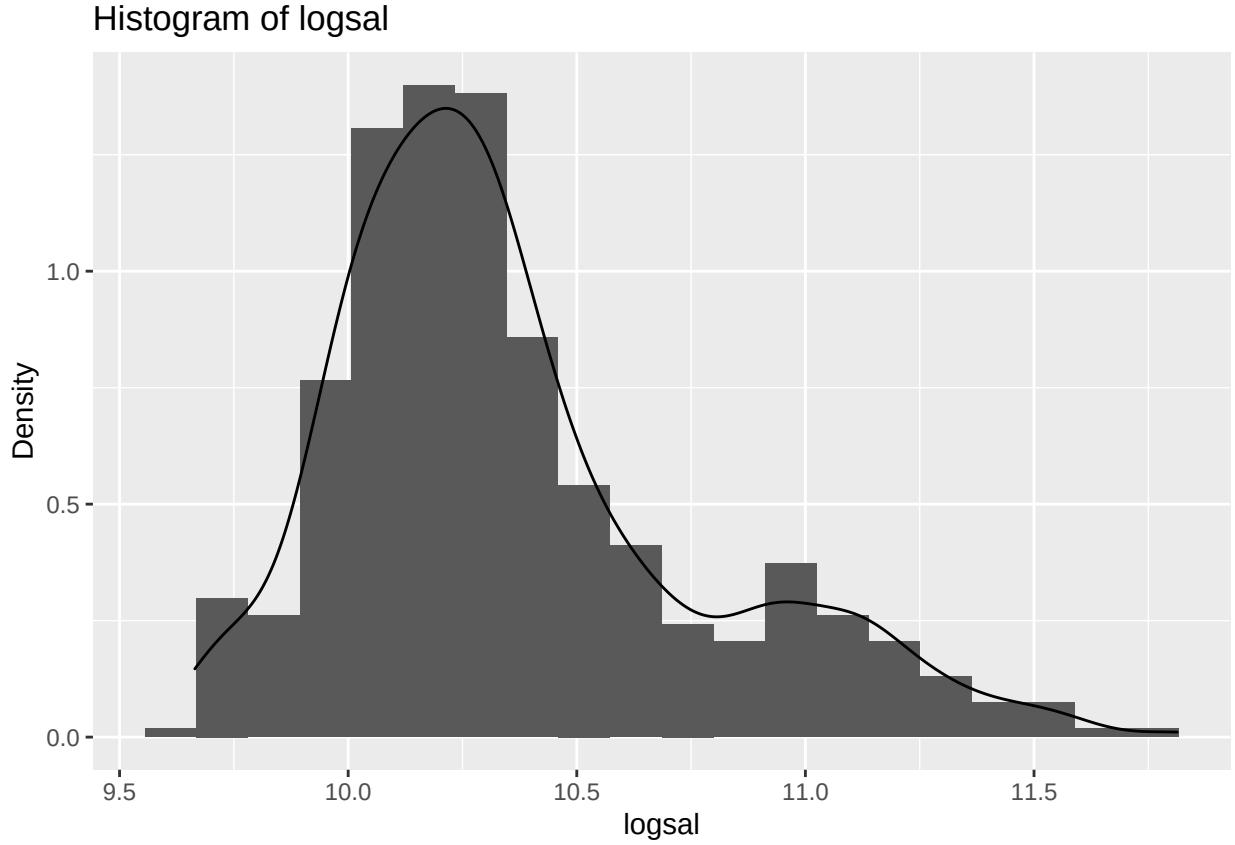
ETC2410 Assignment 1

Alex Wong, Chelaka Paranahewa, Harjot Channa, Jonas Tiong

Question 1

(a)

(i) Histogram of `logsal`



(ii)

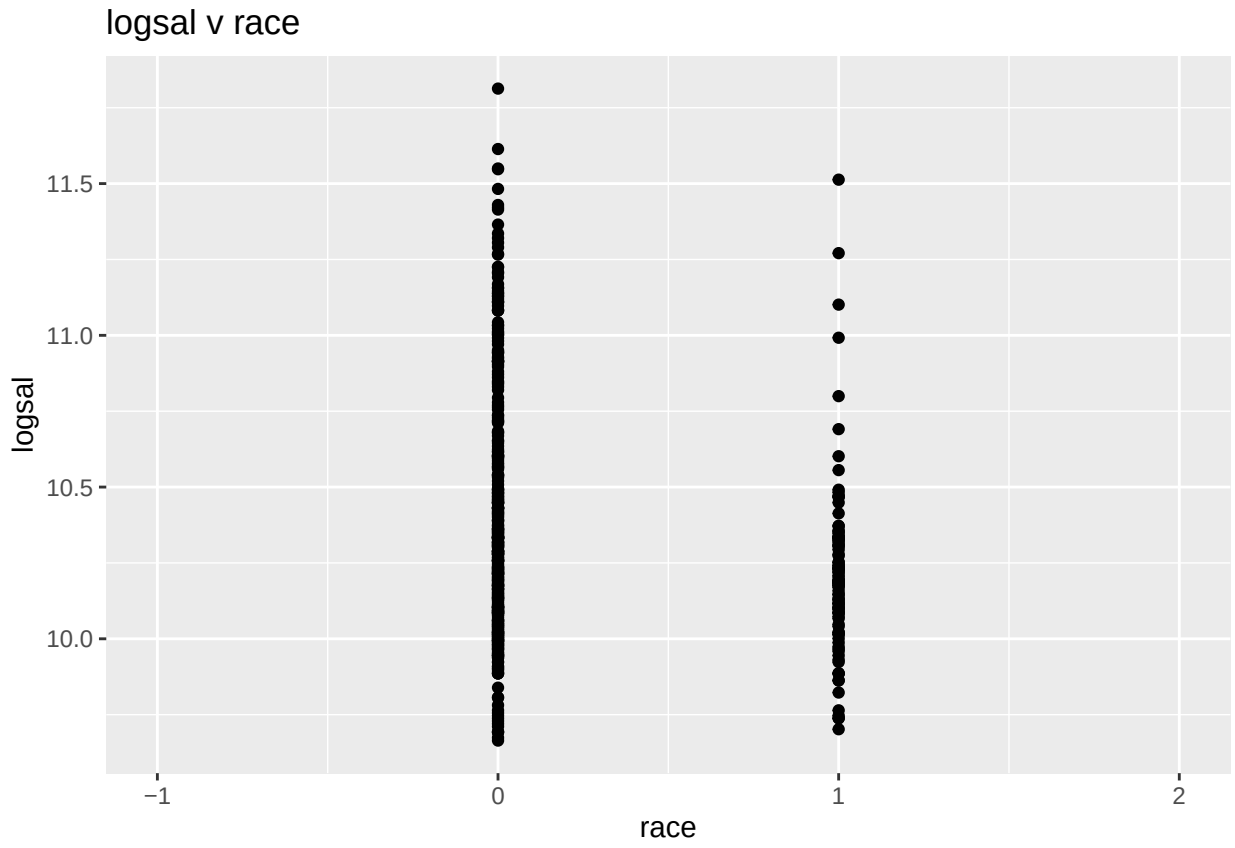
n	mean	median	min	max	sd	skew	kurtosis
474	10.35679	10.27073	9.664596	11.81303	0.3973342	0.994876	0.6471944

As the mean is greater than the median, the distribution is positively skewed indicating that there are high valued outliers (those earning high annual wages). Since the skewness is 0.99, we can say that this distribution is moderately skewed.

On average, an individual in this bank will have a log of annual salary of around 10.36 or, in other words, an annual salary of around $e^{10.36} = 31,470.09$. The maximum annual salary, empirically, is around $e^{11.81} = 135,000.00$ and the lowest salary in the bank is around $e^{9.66} = 15,750.00$ That is a difference of around $119,250.00$.

(b)

(i) Scatterplot of `logsal` and `race`



race：1 if individual i belongs to an ethnic minority, 0 otherwise

logsal：the natural log of individual i's annual salary

(ii) Recall from part (a)ii) that the mean of `logsal` was around 10.36. From the scatterplot it is evident that in this bank, an individual who does not belong to an ethnic minority earns in a wide range of salaries from slightly below to 10.0 to slightly above 11.0 (as per where all the data points

are clustered). In contrast, for an individual who does belong to an ethnic minority this is range is far smaller, somewhere slightly below 10.0 to around 10.5. Hence, the gap from the mean wage of 10.36 is larger for non-minorities which may suggest larger mobility for this category of employees which would be indicative of racial discrimination.

(c)

(i)

$$\widehat{logsal} = \underset{(0.020)}{10.396} - \underset{(0.043)}{0.180}\ race \tag{1}$$

$$R^2 : 0.035414$$
$$n : 474$$

(ii) Our null hypothesis is that race has no effect on an individual's annual (log of their) salary in this bank ($\beta_1 = 0$). The alternative hypothesis is that race has an effect ($\beta_1 \neq 0$).

$$\begin{aligned}
H_0 &: \beta_1 = 0 \\
H_1 &: \beta_1 \neq 0 \\
\text{Significance Level} &: \alpha = 0.05 \\
\text{Test stat and null dist} &: \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-k-1} = t_{472} \\
t_{calc} &= 4.162804 \\
t_{crit} &= 1.9650027 \\
\text{Decision rule} &: \text{reject } H_0 \text{ if } |t_{calc}| > t_{crit} \\
\text{Decision} &: \text{Since 4.162804 > 1.9650027, reject } H_0
\end{aligned}$$

In conclusion, at 0.05 level of significance, we reject the null hypothesis that race has no effect on (the log of) an individual's annual salary in favour of the alternative hypothesis race has an effect.

(iii) $\hat{\beta}_1$ measures the average difference in the log of an individual's annual salary in the bank (thus proportionate difference) between someone who belongs to an ethnic minority and someone who does not.

Hence, the average difference, according to our model, in an individual's annual salary between someone who belongs to an ethnic minority is $e^{-0.180}$ times the annual salary of someone who does not belong to an ethnic minority. Or, in other words, the average difference of the log annual salary is that someone who does not belong to an ethnic minority earns -0.180 more than someone who is part of ethnic minority.

(iv) This model does not provide conclusive evidence of racial discrimination in salaries paid by the bank. This is because it does not account for (condition on) confounding variables - variables that causally effect an individual's annual salary and whether they belong to an ethnic minority. For example, an individual's level of education may be a variable of interest; the individual may be an immigrant to which US' immigration policy and its skill stream for accepting immigrants (based on education) may decide whether or not someone (in America) belongs to an ethnic minority. And, for an individual's annual salary, may be effected by if the company values (and thus willing to pay wages that are higher) for someone based on the education that they have.

(v) As the coefficient of determination is around $R^2 = 0.035$, this means that around 3.5 per cent of the sample variation in the log of an individual's annual salary is explained by race in this bank. Hence, this model is a very low fit for describing what variables effect an individual's annual salary, where the remaining 96.5 per cent may be caused be unaccounted for variables like education (as aforementioned) or inherent variability.

```
## [1] 0.03541368
```

(vi) β_0 measures the conditional mean of the log of an individual's annual salary in this bank who does not belong to an ethnic minority. This value is 10.396. In other words, the conditional mean of an individual's annual salary who is not part of of ethnic minority is $e^{10.396}$.

(vii)

$$\begin{aligned}\text{Confidence interval} &: (\hat{\beta}_0 - t_{472@0.025}se(\hat{\beta}_0), \hat{\beta}_0 + t_{472@0.025}se(\hat{\beta}_0)) \\ &: (10.3963932 - 1.9650027 \times 0.0203088, 10.3963932 + 1.9650027 \times 0.0203088) \\ &: (10.3564863, 10.4363001)\end{aligned}$$

(viii) We are 95 per cent confident, on average, that the population mean of the the log of an individual's annual salary in this bank for someone who does not belong to an ethnic minority is between 10.356 and 10.436. That is, 95 per cent of the time, the interval between $e^{10.356}$ and $e^{10.436}$ will contain the population mean of an individual's annual salary in this bank for someone who does not belong to a ethnic minority.

(ix) Since $\beta_0 = 0$, in a hypothesis test of individual significance at 0.05, does not fall within the 95 per cent confidence interval $10.356, 10.436$ we conclude that $\hat{\beta}_0$ is statistically significant. That is, we reject the null hypothesis that $\beta_0 = 0$ for the alternative hypothesis that $\beta_0 \neq 0$.

(d)

(i)

$$\widehat{logsal} = \underset{(0.390)}{4.026} + \underset{(0.004)}{0.024}\, educ + \underset{(0.045)}{0.601}\, logssal + \underset{(0.019)}{0.061}\, gender - \underset{(0.019)}{0.043}\, race + \underset{(0.016)}{0.121}\, jobcat \tag{2}$$

$$R^2 : 0.826101$$
$$n : 474$$

(ii) A regressor is individually insignificant at 0.01 level of significance if its p-value is larger than 0.01, p-value > 0.01. So according to our model only race is individually insignificant at this level as p-value$_{race} = 0.027 > 0.01$.

(iii)

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$
$$H_1 : \underset{i\in\{1,2,3,4,5\}}{\exists} \beta_i \neq 0 \text{at least one regressor coef is zero}$$
$$\text{Significance Level} : \alpha = 0.05$$
$$\text{Unresticted Model} : \widehat{logsal} = \underset{(0.390)}{4.026} + \underset{(0.004)}{0.024}educ + \underset{(0.045)}{0.601}logssal + \underset{(0.019)}{0.061}gender - \underset{(0.019)}{0.043}race + \underset{(0.016)}{0.121}jobcat$$
$$\text{Resticted Model} : \widehat{logsal} = \underset{(0.018)}{10.357}$$
$$\text{Test stat and null dist} : \frac{(SSR_R - SSR_{UR})}{SSR_{UR}} \frac{(n-k-1)}{q} \sim F_{(q,n-k-1)} = F_{5,468}$$
$$F_{calc} = 444.6423804$$
$$F_{crit} = 2.5936131$$
$$\text{Decision rule} : \text{reject } H_0 \text{ if } t_{calc} > F_{crit}$$
$$\text{Decision} : \text{Since } 444.6423804 > 2.5936131, \text{ reject } H_0$$

In conclusion, at 0.05 level of significance, we reject the null hypothesis that an individual's number of years of education, their gender, the job category within the bank, their log of their starting annual salary, and if they belong to an ethnic minority are jointly insignificant in effecting (the log of an) individual's annual salary in favour of the alternative hypothesis that at least one of these variables are significant.

(iv) $\hat{\beta}_4$ measures the average difference in the log of an individual's annual salary in the bank (thus proportionate difference) between someone who belongs to an ethnic minority and someone who does not, controlling for the the number of years of education, the gender, the job category within the bank, and the log of an individual's starting salary.

Hence *this* average difference is -0.043; someone who does not belong to an ethnic minority, ceteris paribus, earns 0.043 more in annual log salary than someone does. In other words, for an individual who belongs to an ethnic minority, he/she earns $e^{-0.043}$ times the annual salary of someone who does not belong an ethnic minority, on average, ceteris paribus.

(v) $\hat{\beta}_3 - \hat{\beta}_4 = 0.103$. This means that an individual in the bank who is male and does not belong to an ethnic minority, on average controlling for education, their starting salary, and job category, earns 0.103 more than a female and does belong to an ethnic minority in (the log of) their annual salary. In other words, a male who does not belong to an ethnic minority earns $e^{0.103}$ times their counterpart (female and belongs to an ethnic minority) annually in salaries, on average, controlling for education, their starting salary, and job category.

(vi) It is likely that this model does provide conclusive evidence of racial discrimination in salaries paid by the bank, as per the individual significance of the regression coefficient for `race` and the joint significance of all regressors. This is because the model takes into account possible confounding variables to `race` and `logsal` — years of education, one's gender, the job category, and the (log of) starting salary — which would allow us to make causal statements about an individual's racial status on the (log of their) annual salary. In other words, we are unlikely to have committed **omitted variable bias** in this model by our specification of variables.

(vii) The new model has $R^2 = 0.826$, meaning around 82.6 per cent of the sample variation in the log of an individual's annual salary is explained by the independent variables. This is a much better comparison to the old model where only 3.5 per cent was explained by the independent variables, leaving much variation unexplained (about a $(0.826 - 0.035) \times 100\% = 79.1\%$ difference).

Moreover, given that the variance inflation factor (VIF) is not more than 10, which would be indicative of high correlation between regressors, but rather very small at below 2 for each save `logssal` (at around 4.21), this suggests that our new model has not overspecified and that this R^2 is not high largely because of adding more variables.

Model	R-Squared
Old	0.0354137
New	0.8261007

Regressor	VIF
educ	1.892869
logssal	4.205978
gender	1.505895
race	1.077052
jobcat	2.517574

(e)

We are working with the same model prior:

$$\widehat{logsal} = \underset{(0.390)}{4.026} + \underset{(0.004)}{0.024}\, educ + \underset{(0.045)}{0.601}\, logssal + \underset{(0.019)}{0.061}\, gender - \underset{(0.019)}{0.043}\, race + \underset{(0.016)}{0.121}\, jobcat \tag{3}$$

$$R^2 : 0.826101$$
$$n : 474$$

Our null hypothesis is that an individual's gender and race (whether they belong to an ethnic minority) has no effect on the annual (log of their) salary in this bank ($\beta_3 = \beta_4 = 0$). The alternative hypothesis is that it does have an effect (β_3 and/or β_4 is non-zero).

$$H_0 : \beta_3 = \beta_4 = 0$$
$$H_1 : \underset{i \in \{3,4\}}{\exists} \beta_i \neq 0 (\beta_3 \text{ and/or} \beta_4 \text{is non zero})$$
$$\text{Significance Level} : \alpha = 0.1$$
$$\text{Unresticted Model} : \widehat{logsal} = \underset{(0.390)}{4.026} + \underset{(0.004)}{0.024} educ + \underset{(0.045)}{0.601} logssal + \underset{(0.019)}{0.061} gender - \underset{(0.019)}{0.043} race + \underset{(0.016)}{0.121} jobcat$$
$$\text{Resticted Model} : \widehat{logsal} = \underset{(0.356)}{3.439} + \underset{(0.004)}{0.0241} educ + \underset{(0.041)}{0.665} logssal + \underset{(0.016)}{0.117} jobcat$$
$$\text{Test stat and null dist} : \frac{(SSR_R - SSR_{UR})}{SSR_{UR}} \frac{(n-k-1)}{q} \sim F(q, n-k-1) = F(2, 468)$$
$$F_{calc} = 6.4746026$$
$$F_{crit} = 3.0149905$$
$$\text{Decision rule} : \text{reject } H_0 \text{ if } F_{calc} > F_{crit}$$
$$\text{Decision} : \text{Since } 6.4746026 > 3.0149905, \text{ reject } H_0$$

In conclusion, at 0.05 level of significance, we reject the null hypothesis that an individual's gender and whether they belong to ethnic minority has no effect on the (log of their) annual salary in favour of the alternative hypothesis that their gender and/or their race does effect the individual's (log of their) annual salary.

(f)

(i) Given that

$$\mathbb{E}[logsal \mid educ, logssal, gender, race, jobcat] = \beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_3\, gender + \beta_4\, race + \beta_5\, jobcat$$

then our estimated conditional mean has form:

$$\hat{\beta}_0 + \hat{\beta}_1\, educ + \hat{\beta}_2\, logssal + \hat{\beta}_3\, gender + \hat{\beta}_4\, race + \hat{\beta}_5\, jobcat$$

For population A, i.e. for the population of female managers with 12 years of education who belong to a racial minority and received a given starting salary, the average `logsal` is:

$$\begin{aligned}\mathbb{E}[logsal \mid educ = 12, gender = 0, race = 1, jobcat = 3, logssal] &= \beta_0 + \beta_1 12 + \beta_2\, logssal + \beta_4 + \beta_5 3 \\ &= (\beta_0 + 12\beta_1 + \beta_4 + 3\beta_5) + \beta_2\, logssal\end{aligned}$$

$$\mathbb{E}[logsal \mid educ = 12, gender = 0, race = 1, jobcat = 3, logssal] = (\beta_0 + 12\beta_1 + \beta_4 + 3\beta_5) + \beta_2\, logssal \tag{4}$$

(ii) Likewise, for population B, i.e. for the population of male managers with 11 years of education who are not members of a ethnic minority and receives the same starting salary as the individuals in population A, the average `logsal` is:

$$\begin{aligned}\mathbb{E}[logsal \mid educ = 11, gender = 1, race = 0, jobcat = 3, logssal] &= \beta_0 + \beta_1 11 + \beta_2\, logssal + \beta_3 + \beta_5 3 \\ &= (\beta_0 + 11\beta_1 + \beta_3 + 3\beta_5) + \beta_2\, logssal\end{aligned}$$

$$\mathbb{E}[logsal \mid educ = 11, gender = 1, race = 0, jobcat = 3, logssal] = (\beta_0 + 11\beta_1 + \beta_3 + 3\beta_5) + \beta_2\, logssal \tag{5}$$

(iii) Given the null hypothesis that the average `logsal` of population A is equal to that of population B

$$\begin{aligned}&\mathbb{E}[logsal \mid educ = 12, gender = 0, race = 1, jobcat = 3, logssal] = \\ &\mathbb{E}[logsal \mid educ = 11, gender = 1, race = 0, jobcat = 3, logssal]\end{aligned}$$

, it follows that:

$$\begin{aligned}&\mathbb{E}[logsal \mid educ = 12, gender = 0, race = 1, jobcat = 3, logssal] \\ -&\mathbb{E}[logsal \mid educ = 11, gender = 1, race = 0, jobcat = 3, logssal] = 0\end{aligned}$$

$$(\beta_0 + 12\beta_1 + \beta_4 + 3\beta_5) + \beta_2\, logssal - (\beta_0 + 11\beta_1 + \beta_3 + 3\beta_5) - \beta_2\, logssal = \beta_1 + \beta_4 - \beta_3 = 0$$

So the restriction that follows from the null hypothesis is that $\beta_1 + \beta_4 = \beta_3$. The negation of this statement is that $\beta_1 + \beta_4 \neq \beta_3$ which is the alternative hypothesis, where the average `logsal` of population A is not equal to that of population B.

(iv) Under the null hypothesis, $\beta_1 + \beta_4 = \beta_3$. So, rearranging to get $\beta_1 = \beta_3 - \beta_4$ we can substitute this into our unrestricted model (the population model):

$$\mathbb{E}[logsal \mid educ, logssal, gender, race, jobcat] = \beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_3\, gender + \beta_4\, race + \beta_5\, jobcat$$

to get:

$$\begin{aligned} &\beta_0 + (\beta_3 - \beta_4)\, educ + \beta_2\, logssal + \beta_3\, gender + \beta_4\, race + \beta_5\, jobcat \\ &= \beta_0 + \beta_3\, educ - \beta_4\, educ + \beta_2\, logssal + \beta_3\, gender + \beta_4\, race + \beta_5\, jobcat \\ &= \beta_0 + \beta_2\, logssal + \beta_3(educ + \ gender) + \beta_4(race \ - \ educ) + \beta_5\, jobcat \end{aligned}$$

let $U = educ + gender$ and $V = race - educ$, so

$$\beta_0 + \beta_2\, logssal + \beta_3 U + \beta_4 V + \beta_5\, jobcat \tag{6}$$

This expression is our restricted model. We will use these two models for our F-test when we come to test the null hypothesis that $\beta_1 + \beta_4 = \beta_3$ (tantamount to holding that the two populations A and B have the same `logsal`) against the alternative hypothesis that $\beta_1 + \beta_4 \neq \beta_3$ (that the two populations A and B do not have the same average `logsal`).

(v) So our restricted model is:

$$\widehat{logsal} = \underset{(0.371)}{3.685} + \underset{(0.043)}{0.636}\, logssal + \underset{(0.012)}{0.023} U - \underset{(0.012)}{0.003} V + \underset{(0.016)}{0.119} jobcat \tag{7}$$

(vi)

$$
\begin{aligned}
H_0 &: \beta_1 + \beta_4 = \beta_3 \\
H_1 &: \beta_1 + \beta_4 \neq \beta_3 \\
\text{Significance Level} &: \alpha = 0.1 \\
U &: educ + gender \\
V &: race - educ \\
\text{Unresticted Model} &: \widehat{logsal} = \underset{(0.390)}{4.026} + \underset{(0.004)}{0.024} educ + \underset{(0.045)}{0.601} logssal + \underset{(0.019)}{0.061} gender \\
&\quad - \underset{(0.019)}{0.043} race + \underset{(0.016)}{0.121} jobcat \\
\text{Resticted Model} &: \widehat{logsal} = \underset{(0.370848)}{3.684869} + \underset{(0.042768)}{0.635941} logssal + \underset{(0.012260)}{0.022563} U - \underset{(0.012137)}{0.002899} V \\
&\quad + \underset{(0.015798)}{0.118732} jobcat \\
\text{Test stat and null dist} &: \frac{(SSR_R - SSR_{UR})}{SSR_{UR}} \frac{(n-k-1)}{q} \sim F(q, n-k-1) = F(2, 468) \\
F_{calc} &= 7.0807365 \\
F_{crit} &= 3.8614052 \\
\text{Decision rule} &: \text{reject } H_0 \text{ if } F_{calc} > F_{crit} \\
\text{Decision} &: \text{Since } 7.0807365 > 3.8614052, \text{ reject } H_0
\end{aligned}
$$

In conclusion, at 0.05 level of significance, we reject the null hypothesis that the average of the `logsalary` for population A is the same population B in favour of the alternative hypothesis that they are not equal. In other words, there is a difference between an individual who is female, a manager, with 12 years of education, belongs to an ethnic minority, and has a starting salary *and* an individual who is male, a manager, has 11 years of education, is not an member of a ethnic minority, and also has a starting salary.

(g)

(i) By definition, the racial pay gap for males is:

$$
\begin{aligned}
&\mathbb{E}[logsal \mid gender = 1, race = 1, educ, logssal, jobcat] \\
-&\mathbb{E}[logsal \mid gender = 1, race = 0, educ, logssal, jobcat]
\end{aligned}
$$

Given

$$
\mathbb{E}[logsal \mid educ, logssal, gender, race, jobcat] = \beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_3\, gender + \beta_4\, race + \beta_5\, jobcat
$$

it follows that:

$$\begin{aligned}
&\mathbb{E}[logsal \mid gender = 1, race = 1, educ, logssal, jobcat] \\
-&\mathbb{E}[logsal \mid gender = 1, race = 0, educ, logssal, jobcat] \\
&= \beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_3 + \beta_4 + \beta_5\, jobcat \\
&\quad -(\beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_3 + \beta_5\, jobcat) \\
&= \beta_4
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}[logsal \mid gender = 1, race = 1, educ, logssal, jobcat] - \mathbb{E}[logsal \mid gender = 1, \\
&\qquad race = 0, educ, logssal, jobcat] = \beta_4 \quad (8)
\end{aligned}$$

So the racial pay gap for males is β_4.

(ii) By definition, the racial pay gap for females is:

$$\begin{aligned}
&\mathbb{E}[logsal \mid gender = 0, race = 1, educ, logssal, jobcat] \\
-&\mathbb{E}[logsal \mid gender = 0, race = 0, educ, logssal, jobcat
\end{aligned}$$

Given

$$\begin{aligned}
\mathbb{E}[logsal \mid educ, logssal, gender, race, jobcat] = \beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_3\, gender + \beta_4\, race \\
+\beta_5\, jobcat
\end{aligned}$$

it follows that:

$$\begin{aligned}
&\mathbb{E}[logsal \mid gender = 0, race = 1, educ, logssal, jobcat] \\
-&\mathbb{E}[logsal \mid gender = 0, race = 0, educ, logssal, jobcat] \\
&= \beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_4 + \beta_5\, jobcat \\
&\quad -(\beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_5\, jobcat) \\
&= \beta_4
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}[logsal \mid gender = 0, race = 1, educ, logssal, jobcat] \\
&\qquad - \mathbb{E}[logsal \mid gender = 0, race = 0, educ, logssal, jobcat] = \beta_4 \quad (9)
\end{aligned}$$

So the racial pay gap for females is also β_4.

These results make sense, for another interpretation of β_4 is the partial effect of being in an ethnic minority on the log of an individual's starting salary against not being in an ethnic minority, that is, holding all else constant. Hence, whether or not we are looking at male or female racial pay gap is already accounted for in this model by β_4.

(h)

(i) To allow in our model for the racial pay gap to vary by gender, we may introduce the interaction term $race \times gender$, allowing `race` to vary as a function of `gender`. Thus,

$$\mathbb{E}[logsal \mid educ, logssal, gender, race, jobcat] = \beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_3\, gender + \beta_4\, race + \beta_5\, jobcat + \beta_6\, race \times gender \quad (10)$$

(ii) The racial pay gap for males, controlling for education, starting salary, and job category is:

$$\begin{aligned} &\mathbb{E}[logsal \mid gender = 1, race = 1, educ, logssal, jobcat] \\ -&\mathbb{E}[logsal \mid gender = 1, race = 0, educ, logssal, jobcat] \end{aligned}$$

According to our new model, this is

$$\begin{aligned} &\mathbb{E}[logsal \mid gender = 1, race = 1, educ, logssal, jobcat] \\ -&\mathbb{E}[logsal \mid gender = 1, race = 0, educ, logssal, jobcat] \\ =\ &\beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_3 + \beta_4 + \beta_5\, jobcat + \beta_6 \\ &-(\beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_3 + \beta_5\, jobcat) \\ =\ &\beta_4 + \beta_6 \end{aligned}$$

$$\begin{aligned} &\mathbb{E}[logsal \mid gender = 1, race = 1, educ, logssal, jobcat] \\ &- \mathbb{E}[logsal \mid gender = 1, race = 0, educ, logssal, jobcat] = \beta_4 + \beta_6 \end{aligned} \quad (11)$$

So *this* racial pay gap for males is equal to $\beta_4 + \beta_6$.

(iii) The racial pay grap for females, controlling for education, starting salary, and job category is:

$$\begin{aligned} &\mathbb{E}[logsal \mid gender = 0, race = 1, educ, logssal, jobcat] \\ -&\mathbb{E}[logsal \mid gender = 0, race = 0, educ, logssal, jobcat] \end{aligned}$$

According to our new model, this is

$$\begin{aligned} &\mathbb{E}[logsal \mid gender = 0, race = 1, educ, logssal, jobcat] \\ -&\mathbb{E}[logsal \mid gender = 0, race = 0, educ, logssal, jobcat] \\ =\ &\beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_4 + \beta_5\, jobcat \\ &-(\beta_0 + \beta_1\, educ + \beta_2\, logssal + \beta_5\, jobcat) \\ =\ &\beta_4 \end{aligned}$$

$$\mathbb{E}[logsal \mid gender = 0, race = 1, educ, logssal, jobcat] - \mathbb{E}[logsal \mid gender = 0, race = 0, educ, logssal, jobcat] = \beta_4 \quad (12)$$

So *this* racial pay gap for females is equal to β_4.

(iv) The null hypothesis is that, controlling for education, starting salary, and job category, the racial pay gap for males and females is the same. Hence, from part (h)ii) and part (h)iii) this statement is equivalent to the stating that $\beta_4 + \beta_6 = \beta_4$ which in other words is to say that $\beta_6 = 0$.

$$\begin{aligned}
H_0 &: \beta_6 = 0 \\
H_1 &: \beta_6 \neq 0 \\
\text{Significance Level} &: \alpha = 0.1 \\
\text{Est Reg} &: \underset{(0.390)}{4.026} + \underset{(0.004)}{0.024}educ + \underset{(0.045)}{0.601}logssal + \underset{(0.019)}{0.061}gender - \underset{(0.019)}{0.043}race + \underset{(0.016)}{0.121}jobcat \\
\text{Test stat and null dist} &: \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-k-1} = t_{467} \\
t_{calc} &= 0.4901646 \\
t_{crit} &= 1.648123 \\
\text{Decision rule} &: \text{reject } H_0 \text{ if } |t_{calc}| > t_{crit} \\
\text{Decision} &: \text{Since } 0.4901646 > 1.648123, \text{ reject } H_0
\end{aligned}$$

In conclusion, at 0.10 level of significance, we reject the null hypothesis that, controlling for education, starting salary, and job category, the racial pay gap for males and females is the same in favour of the alternative hypothesis that they are not.

Question 2

(a)

$$
\begin{aligned}
&\hat{u}_i = residuals \\
&\hat{u}_i = y_i - \hat{y}_i,\ i = 1, 2, ...n && \text{from Assignment Property (7)} \\
&\quad = y_i - \left(\hat{\beta}_o + \hat{\beta}_1 x_i\right) && \text{by definition } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \\
&\quad = y_i - \hat{\beta}_o - \hat{\beta}_1 x_i
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{n} \hat{u}_i &= \sum_{i=1}^{n} y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_0 - \sum_{i=1}^{n} \hat{\beta}_1 x_i \\
&= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i - n\hat{\beta}_0 && \text{rearranging terms and summing } \hat{\beta}_0 \\
&= \frac{n}{n} \sum_{i=1}^{n} y_i - \frac{n}{n} \sum_{i=1}^{n} \hat{\beta}_1 x_i - n\hat{\beta}_0 && \text{times first and second sum by } \frac{n}{n} \\
&= n\bar{y} - n\hat{\beta}_i \bar{x} - n\hat{\beta}_0 \\
&= n\bar{y} - n\hat{\beta}_i \bar{x} - n(\bar{y} - \hat{\beta}_i \bar{x}) && \text{substituting from Assignment Property (6)} \\
&= n\bar{y} - n\hat{\beta}_i \bar{x} - n\bar{y} + n\hat{\beta}_i \bar{x} \\
&= n\bar{y} - n\bar{y} - n\hat{\beta}_i \bar{x} + n\hat{\beta}_i \bar{x} \\
&= 0
\end{aligned}
$$

(b)

$$SSR\ (b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

$$\left.\frac{\partial SSR\,(b_0, b_1)}{\partial b_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

$$\left.\frac{\partial SSR\,(b_0, b_1)}{\partial b_i}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2\sum_{i=1}^{n} x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \quad \text{call this (1)}$$

$$\begin{aligned} \hat{u}_i &= y_i - \hat{y}_i,\ i = 1, 2, ...n && \text{from Assignment Property (7)} \\ &= y_i - \left(\hat{\beta}_o + \hat{\beta}_1 x_i\right) && \text{by definition } \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= y_i - \hat{\beta}_o - \hat{\beta}_1 x_i && \text{call this (2)} \end{aligned}$$

$$-2\sum_{i=1}^{n} x_i(\hat{u}_i) = 0 \qquad \text{substituting (2) into (1)}$$

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0 \qquad \text{call this (3)}$$

Now if $x = \underset{(n\times 1)}{\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}}$ and $\hat{u} = \underset{(n\times 1)}{\begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{pmatrix}}$ then the dot product is

$$\begin{aligned} x'\hat{u} &= (x_1 \quad ... \quad x_n) \begin{pmatrix} \hat{u}_1 \\ \vdots \\ \hat{u}_n \end{pmatrix} \\ &= x_1\hat{u}_1 + \ ... \ + x_n\hat{u}_n && \text{just the linear combinations of column vector } u \\ & && \text{where the scalars are the components of } x \\ &= \sum_{i=1}^{n} x_i \hat{u}_i \\ &= 0 && \text{from (3)} \end{aligned}$$

Question 3

(a) $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$ for $i=1,\dots,n$ from question

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i - (x_i-\bar{x})\bar{y}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i - \sum_{i=1}^{n}(x_i-\bar{x})\bar{y}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i + \sum_{i=1}^{n}(\bar{x}\bar{y}-x_i\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i + \sum_{i=1}^{n}\bar{x}\bar{y} - \bar{y}\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i + n\bar{x}\bar{y} - \bar{y}n\sum_{i=1}^{n}\frac{x_i}{n}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i + n\bar{x}\bar{y} - \bar{y}n\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

(b) (i) $\mathbb{E}[\hat{\beta}_1 \mid x_1, \cdots, x_n]$

$= \mathbb{E}[\hat{\beta}_1 \mid X]$, $X = (x_1, \cdots, x_n)$

$= \mathbb{E}\left[\dfrac{\sum_{i=1}^{n} (x_i - \bar{x}) y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \mid X \right]$, $i = 1, \cdots, n$ from (9)

$= \mathbb{E}\left[\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \mid X \right]$ from (4)

$= \mathbb{E}\left[\dfrac{\sum_{i=1}^{n} [(x_i - \bar{x})\beta_0 + (x_i - \bar{x})\beta_1 x_i + (x_i - \bar{x}) u_i]}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \mid X \right]$

$= \mathbb{E}\left[\dfrac{\beta_0 \sum_{i=1}^{n} (x_i - \bar{x}) + \beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i + \sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \mid X \right]$

Now, $\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x}$

$= n \sum_{i=1}^{n} \frac{x_i}{n} - n\bar{x}$

$= n\bar{x} - n\bar{x}$

$= 0$

So, $\mathbb{E}[\hat{\beta}_1 \mid x_1, \cdots, x_n] = \mathbb{E}\left[\dfrac{\beta_1 \sum_{i=1}^{n} (x_i - \bar{x}) x_i + \sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \mid X \right]$

$$= \mathbb{E}\left[\beta_1 \frac{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \,\middle|\, X \right]$$

Now $$\frac{\sum_{i=1}^{n} (x_i - \bar{x}) x_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) x_i - \bar{x} \sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$ as

$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ from before.

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$= 1$$

So, $$\mathbb{E}[\hat{\beta}_1 \mid x_1, \dots, x_n] = \mathbb{E}\left[\beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \,\middle|\, X\right]$$

$$= \mathbb{E}[\beta_1 \mid X] + \mathbb{E}\left[\frac{\sum_{i=1}^{n} (x_i - \bar{x}) u_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \,\middle|\, X\right]$$

$$= \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \mathbb{E}[u_i \mid X]$$

$= \beta_1$ as $\mathbb{E}[u_i \mid X] = \mathbb{E}[u_i \mid x_1, \dots, x_n]$

$= 0$

from zero conditional mean assumption.

(ii) $\mathrm{Var}\left[\hat{\beta} \mid x_1, \ldots, x_n\right] = \mathrm{Var}\left[\hat{\beta} \mid X\right]$, $X = (x_1, \ldots, x_n)$

$$= \mathrm{Var}\left[\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \mid X\right] \quad \text{from part (b) ii)}$$

$$= \mathrm{Var}\left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})u_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \mid X\right] \quad \text{as } \beta_1 \text{ is a constant}$$

$$= \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2 \mathrm{Var}\left[u_i \mid X\right] \quad \text{as only } u_i \text{ is a r.v.}$$

Now, $\mathrm{Var}\left[y_i \mid x_1, \ldots, x_n\right] = \mathrm{Var}\left[y_i \mid X\right]$

$$= \mathrm{Var}\left[\beta_0 + \beta_1 x_i + u_i \mid X\right] \quad \text{from (4)}$$

$$= \mathrm{Var}\left[u_i \mid X\right] \quad \text{as only } u_i \text{ is a r.v.}$$

$$= \sigma^2 \quad \text{from (8)}$$

So, $\mathrm{Var}\left[\hat{\beta} \mid x_1, \ldots, x_n\right] = \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^2 \sigma^2$

$$= \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\,\sigma^2$$

(iii) $$Var[\hat{\beta}_1] = \mathbb{E}\left[Var[\hat{\beta}_1 | x_1, \dots, x_n]\right] + Var\left[\mathbb{E}[\hat{\beta}_1 | x_1, \dots, x_n]\right]$$

$$= \mathbb{E}\left[\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2\right] + Var[\beta_1]$$

from part (b) i) and part (b) ii)

$$= \mathbb{E}\left[\frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sigma^2\right]$$ as β_1 is a constant

$$= \frac{\sigma^2}{\mathbb{E}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]}$$ as σ^2 is a constant

Now, $$\mathbb{E}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right] = \frac{n-1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= (n-1)\mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]$$

$$= (n-1)\mathbb{E}\left[\hat{\sigma}_x^2\right]$$

So, $$Var[\hat{\beta}_1] = \frac{\sigma^2}{(n-1)\mathbb{E}\left[\hat{\sigma}_x^2\right]}$$

(c) From part (b), we are aware of two things:

$$\mathbb{E}[\hat{\beta}_1 | x_1, \dots, x_n] = \beta_1$$

and $$\text{Var}[\hat{\beta}_1 | x_1, \dots, x_n] = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Now, the circumstances that provide the best opportunity to precisely estimate $\hat{\beta}_1$ is when $\hat{\beta}_1$ is BLUE (best linear unbias estimator). This occurs when we ~~have~~ our population model is linear in its parameters and is correctly specified, data on $(x_1, \dots, x_n)$ represents a random sample from the population described by the model, there is variation in x, we have zero conditional mean $\mathbb{E}[u_i | x_1, \dots, x_n] = 0$, and homoskedasticity $\text{Var}[u_i | x_1, \dots, x_n] = \sigma^2$. These assumptions imply, according to the Gauss-Markov theorem, that $\hat{\beta}_1$ is BLUE. Hence, we know that $\frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ is the lowest variance of any linear estimator of β_1, i.e. we can precisely estimate β_1.

Appendix

Dependent Variable: LOGSAL
Method: Least Squares
Date: 04/10/22 Time: 20:15
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	10.39639	0.020309	511.9146	0.0000
RACE	-0.180486	0.043357	-4.162804	0.0000

R-squared	0.035414	Mean dependent var	10.35679
Adjusted R-squared	0.033370	S.D. dependent var	0.397334
S.E. of regression	0.390648	Akaike info criterion	0.962193
Sum squared resid	72.03011	Schwarz criterion	0.979751
Log likelihood	-226.0397	Hannan-Quinn criter.	0.969098
F-statistic	17.32894	Durbin-Watson stat	1.586616
Prob(F-statistic)	0.000037		

Figure (1) Q1 (c) i) OLS regression for race on logsal

Dependent Variable: LOGSAL
Method: Least Squares
Date: 04/27/22 Time: 18:15
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.025756	0.390102	10.31976	0.0000
EDUC	0.024382	0.003653	6.675102	0.0000
LOGSSAL	0.600624	0.044518	13.49169	0.0000
GENDER	0.060849	0.018852	3.227697	0.0013
RACE	-0.042609	0.019187	-2.220762	0.0268
JOBCAT	0.120894	0.015717	7.691753	0.0000

R-squared	0.826101	Mean dependent var	10.35679
Adjusted R-squared	0.824243	S.D. dependent var	0.397334
S.E. of regression	0.166576	Akaike info criterion	-0.734152
Sum squared resid	12.98587	Schwarz criterion	-0.681479
Log likelihood	179.9941	Hannan-Quinn criter.	-0.713436
F-statistic	444.6424	Durbin-Watson stat	1.796752
Prob(F-statistic)	0.000000		

Figure (2) Q1 (d) i) OLS regression for edu, logssal, gender, race, jobcat on logsal

Wald Test:
Equation: Untitled

Test Statistic	Value	df	Probability
F-statistic	6.474603	(2, 468)	0.0017
Chi-square	12.94921	2	0.0015

Null Hypothesis: C(4) = C(5) =0
Null Hypothesis Summary:

Normalized Restriction (= 0)	Value	Std. Err.
C(4)	0.060849	0.018852
C(5)	-0.042609	0.019187

Restrictions are linear in coefficients.

Figure (3) Q1 (e) F-statistic for hypothesis test

Dependent Variable: LOGSAL
Method: Least Squares
Date: 04/28/22 Time: 18:49
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.684869	0.370848	9.936325	0.0000
LOGSSAL	0.635941	0.042768	14.86947	0.0000
EDUC+GENDER	0.022563	0.012260	1.840351	0.0663
RACE-EDUC	-0.002899	0.012137	-0.238875	0.8113
JOBCAT	0.118732	0.015798	7.515705	0.0000

R-squared	0.823470	Mean dependent var	10.35679
Adjusted R-squared	0.821964	S.D. dependent var	0.397334
S.E. of regression	0.167652	Akaike info criterion	-0.723355
Sum squared resid	13.18234	Schwarz criterion	-0.679461
Log likelihood	176.4352	Hannan-Quinn criter.	-0.706092
F-statistic	546.9415	Durbin-Watson stat	1.799456
Prob(F-statistic)	0.000000		

Figure (4) Q1 (f) v) Estimated restricted regression model

Dependent Variable: LOGSAL
Method: Least Squares
Date: 04/14/22 Time: 19:09
Sample: 1 474
Included observations: 474

Variable	Coefficient	Std. Error	t-Statistic	Prob.
EDUC	0.008838	0.003716	2.378485	0.0178
LOGSSAL	1.057785	0.004844	218.3683	0.0000
GENDER	-0.023166	0.021511	-1.076931	0.2821
RACE	-0.012765	0.032330	-0.394827	0.6932
JOBCAT	0.015787	0.013265	1.190107	0.2346
RACE*GENDER	0.007307	0.042472	0.172030	0.8635

R-squared	0.786542	Mean dependent var	10.35679
Adjusted R-squared	0.784261	S.D. dependent var	0.397334
S.E. of regression	0.184552	Akaike info criterion	-0.529188
Sum squared resid	15.93990	Schwarz criterion	-0.476515
Log likelihood	131.4176	Hannan-Quinn criter.	-0.508472
Durbin-Watson stat	1.832319		

Figure (5) Q1 (h) iv) Estimated regression model