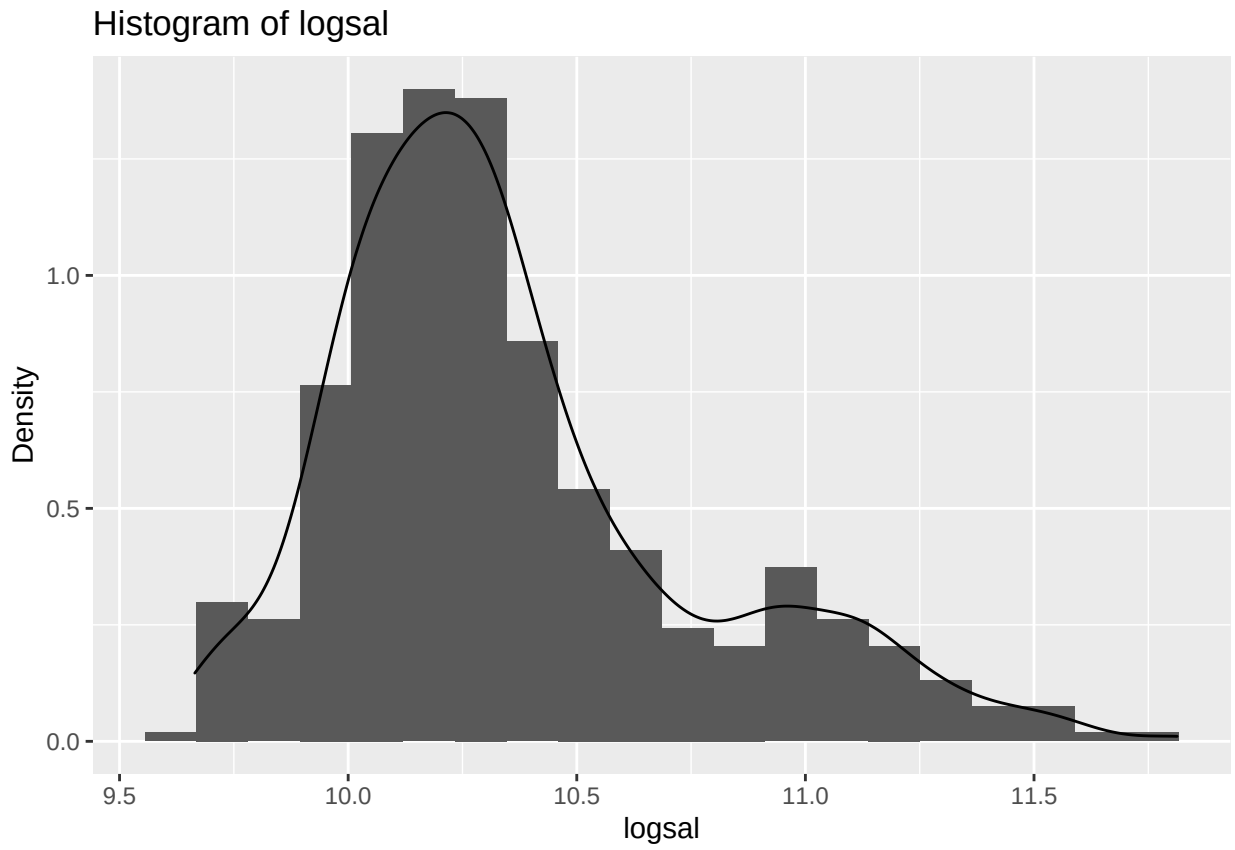
# Group Assignment 1

Alex Wong, Chelaka Paranahewa, Harjot Channa, Jonas Tiong

## Question 1

**Part A**

**Section i** Histogram of `logsal`




**Section ii**

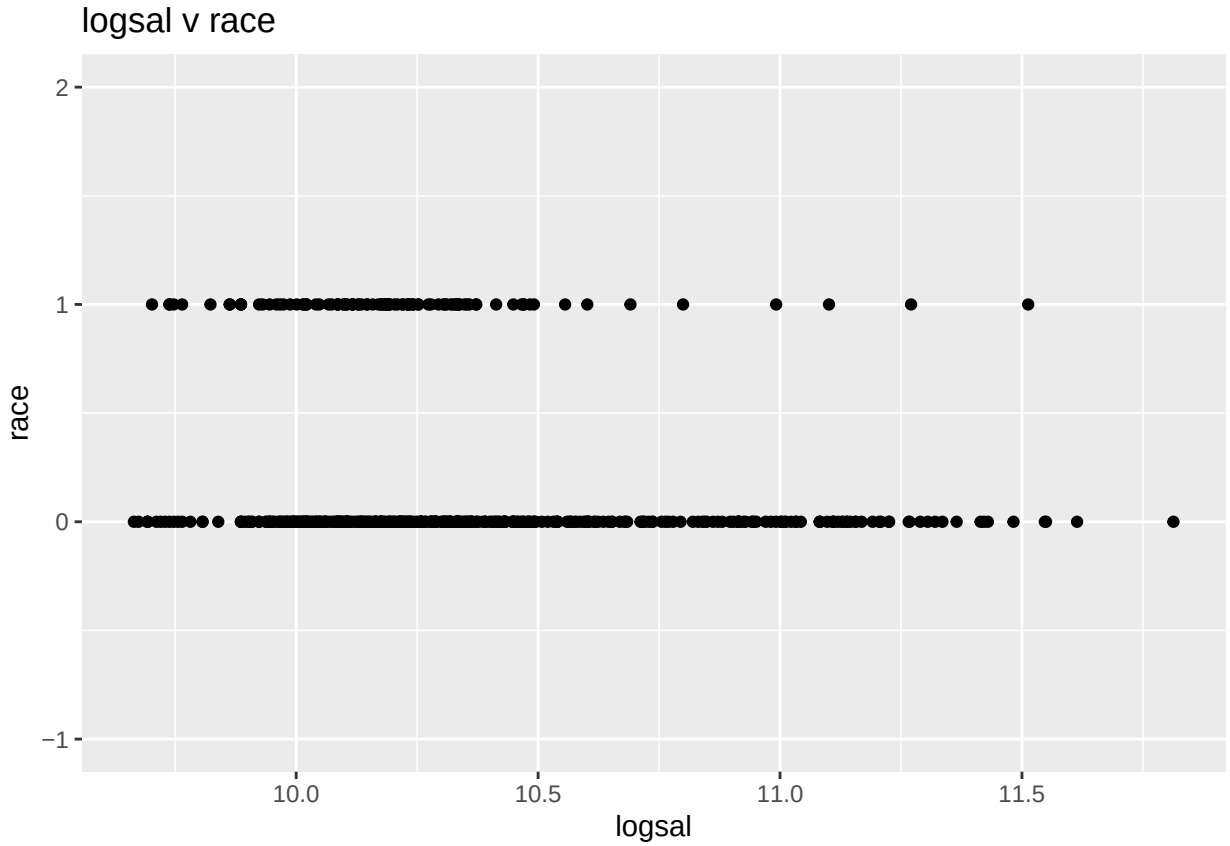
| n | mean | median | min | max | sd | skew | kurtosis |
|---|---|---|---|---|---|---|---|
| 474 | 10.35679 | 10.27073 | 9.664596 | 11.81303 | 0.3973342 | 0.994876 | 0.6471944 |

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 $119,250.00$.

**Part B**

**Section i** Scatterplot of `logsal` and `race`




**Section 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.

**Part C**

**Section i**

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

**Section ii**

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

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.

**Section 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 wage 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.

**Section 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.

**Section 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
```

**Section vi** $\beta_0$ 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 $e^{10.396}$.

**Section vii**

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

**Section viii** We are 95 per cent confident, on average, that the population value 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, the population value of an individual's annual salary in this bank is between $e^{10.356}$ and $e^{10.436}$, on average 95 per cent of the time.

**Section 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.

**Part D**

**Section i**

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

**Section ii** A regressor is individually insignificant at 0.001 level of significance if its p-value is larger than 0.001, p-value $> 0.001$. So, according to our model only gender and race are individually insignificant at this level as p-value$_{gender} = 0.0013 > 0.001$ and p-value$_{race} = 0.027 > 0.001$, respectively.

**Section iii**

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

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.

**Section 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, according to our model, for an individual who belongs to an ethnic minority is $e^{-0.043}$ times the annual salary of someone who does not belong an ethnic minority.

**Section v**

```
## [1] 0.1034585
```

*****MATHSPEAK LATER

$\hat{\beta}_3 - \hat{\beta}_4 = 0.103$ does not have a meaningful interpretation. This is because $\hat{\beta}_3$ is the conditional mean of the log of an individual's annual salary given that they are male and ceteris paribus minus the conditional mean of the log of an individual's annual salary given that they are female and ceteris paribus. $\hat{\beta}_4$ is the conditional mean of the log an individual's annual salary given that they belong to an ethnic minority and ceteris paribus minus the conditional mean of the log of an individual's annual salary given that the do not belong to an ethnic minority and ceteris paribus. Hence, the difference between $\hat{\beta}_3$ and $\hat{\beta}_4$ is the sum of the conditional mean of an individual's annual salary given that they are male, ceteris paribus (including one's racial status), plus the salary given that they are not part of the ethnic minority, ceteris paribus (including one's gender), minus the salary given that they are female, ceteris paribus (including one's racial status), minus the salary given that they are part of an ethnic minority, ceteris paribus (including one's gender). The fact that this sum of different means is conditional on different set of variables makes any descriptive application inapplicable to the bank.

**Section 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.

**Section 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, 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 |

| VIF |
|---|
| 1.892869 |
| 4.205978 |
| 1.505895 |
| 1.077052 |
| 2.517574 |

**Part E**

We are working with the same model prior:

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

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

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.

**Part F**

**Section 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:

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

**Section 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:

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

**Section iii** Given the null hypothesis that the average `logsal` of population A is equal to that of population B $\mathbb{E}[logsal \mid educ = 12, gender = 0, race = 1, jobcat = 3, logssal] = \mathbb{E}[logsal \mid educ = 11, gender = 1, race = 0, jobcat = 3, logssal]$, it follows that:

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

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

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.

**Section 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:

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

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

$$\beta_0 + \beta_2 U + \beta_3\ gender + \beta_4 V + \beta_5\ jobcat$$

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`).

**Section v** So our restricted model is:

$$\widehat{logsal} = \underset{(0.164)}{8.282} - \underset{(0.020)}{0.120}U + \underset{(0.019)}{0.160}gender - \underset{(0.019)}{0.160}V + \underset{(0.015)}{0.227}jobcat$$

**Section vi**

$$\begin{aligned} H_0 &: \beta_3 = \beta_4 = 0 \\ H_1 &: \underset{i \in \{3,4\}}{\exists} \beta_i \neq 0 \\ \text{Significant Level} &: \alpha = 0.1 \\ \text{Unresticted Model} &: \underset{(0.390)}{4.026} + \underset{(0.004)}{0.024}educ + \underset{(0.045)}{0.601}logssal + \underset{(0.019)}{0.061}gender - \underset{(0.019)}{0.043}race + \underset{(0.016)}{0.121}jobcat \\ \text{Resticted Model} &: \widehat{logsal} = \underset{(0.164)}{8.282} - \underset{(0.020)}{0.120}U + \underset{(0.019)}{0.160}gender - \underset{(0.019)}{0.160}V + \underset{(0.015)}{0.227}jobcat \\ \text{Test stat and null dist} &: \frac{(SSR_R - SSR_{UR})}{SSR_{UR}} \frac{(n-k-1)}{q} \sim F(q, n-k-1) = F(2, 468) \\ t_{calc} &= 137.8538771 \\ F_{crit} &= 3.8614052 \\ \text{Decision rule} &: \text{reject } H_0 \text{ if } t_{calc} > F_{crit} \\ \text{Decision} &: \text{Since } 137.8538771 > 3.8614052, \text{ reject } H_0 \end{aligned}$$

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.

**Part G**

**Section i** By definition, the racial pay gap for males is:

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

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:

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

So the racial pay gap for males is $\beta_4$.

**Section ii** By definition, the racial pay gap for females is:

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

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:

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

So the racial pay gap for females is also $\beta_4$.

These results make sense, for another interpretation of $\beta_4$ 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 $\beta_4$.

**Part H**

**Section 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,

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

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

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

According to our new model, this is

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

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

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

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

According to our new model, this is

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

So *this* racial pay gap for females is equal to $\beta_4$.

**Section 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 $\beta_6 = 0$.

$$\begin{aligned}H_0 &: \beta_1 = 0 \\ H_1 &: \beta_1 \neq 0 \\ \text{Significant Level} &: \alpha = 0.1 \\ \text{Est Reg} &: \underset{(0.390)}{4.026} + \underset{(0.004)}{0.024}educ + \underset{(0.045)}{0.601}logssal + \underset{(0.019)}{0.061}gender - \underset{(0.019)}{0.043}race + \underset{(0.016)}{0.121}jobcat \\ \text{Test stat and null dist} &: \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \sim t_{n-k-1} = t_{467} \\ t_{calc} &= 0.4901646 \\ t_{crit} &= 1.648123 \\ \text{Decision rule} &: \text{reject } H_0 \text{ if } |t_{calc}| > t_{crit} \\ \text{Decision} &: \text{Since } 0.4901646 > 1.648123, \text{ reject } H_0\end{aligned}$$

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.