Group Assignment 1

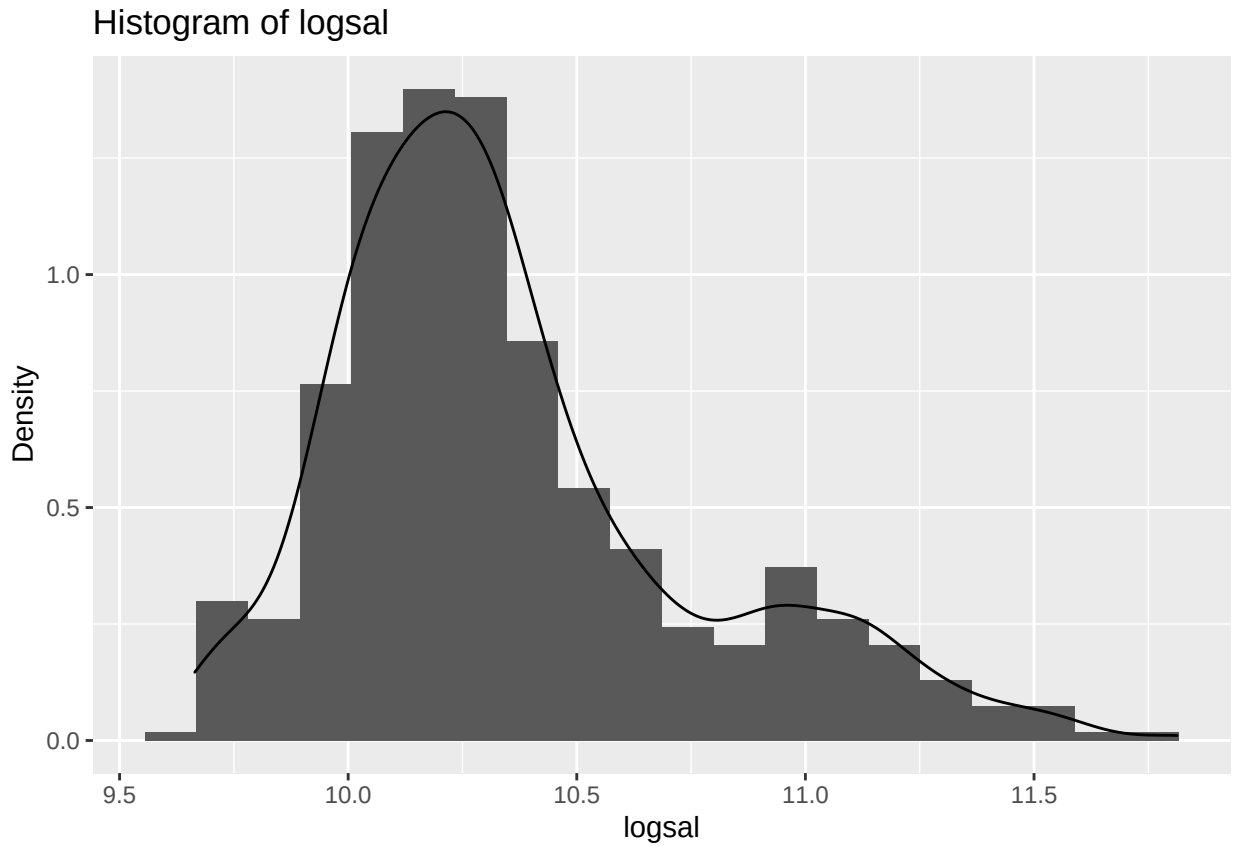
Alex Wong, Chelaka Paranahewa, Harjot Channa, Jonas Tiong

#Question 1

##1.a.

(i) Histogram of logsal

```
dat %>% ggplot(aes(x = logsal)) +
  geom_histogram(aes(y = after_stat(density)),
                 bins = 20) +
  geom_density() +
  labs(
    x = "logsal",
    y = "Density",
    title = "Histogram of logsal"
  )
```



(ii) analysis

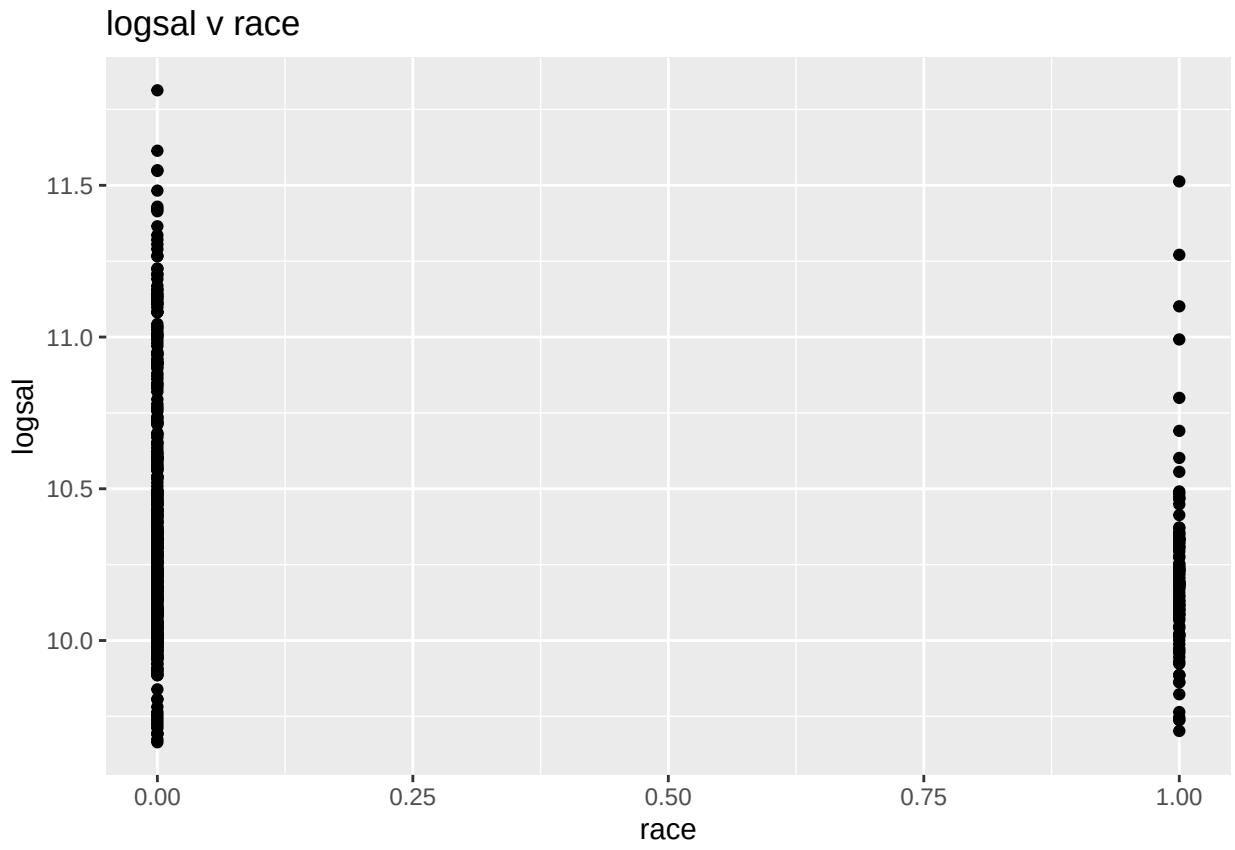
```
logsal_stat <- dat$logsal %>%
  describe() %>%
  as.tibble() %>%
  select(c(n, mean, median, min, max, sd, skew, kurtosis))

kable(logsal_stat)
```

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

##1.b.

(i) scatterplot

```
dat %>% ggplot(aes(x = race, y = logsal)) +
  geom_point() +
  labs(x = "race",
       y = "logsal",
       title = "logsal v race")
```



##1.c.

(i)

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

```r
# OLS regression
fit1 <- lm(logsal ~ race, data = dat)
fit1 %>% summary()
```

```
##
## Call:
## lm(formula = logsal ~ race, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.73180 -0.27303 -0.06764  0.16623  1.41664
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.39639    0.02031 511.915  < 2e-16 ***
## race        -0.18049    0.04336  -4.163 3.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3906 on 472 degrees of freedom
## Multiple R-squared:  0.03541,    Adjusted R-squared:  0.03337
## F-statistic: 17.33 on 1 and 472 DF,  p-value: 3.737e-05
```

(ii)

```r
# level of significance
alpha_1 <- 0.05

# test statistic
tstat1 <- coef(summary(fit1))[2, "Estimate"] /
  coef(summary(fit1))[2, "Std. Error"]
# ~t(472)

# critical value under the null
tcrit1 <- qt(1 - alpha_1/2, 472)

# decision
abs(tstat1) > (tcrit1)
```

```
## [1] TRUE
```

```r
#make something that says reject H0
```

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 that it 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 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.

(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 or not they belong to an ethnic minority. For example, an individual's level of education may 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 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 (being part of a ethnic minority or not) 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.

```
SSR1 <- sum(resid(fit1)^2)
SST1 <- sum((dat$logsal - mean(dat$logsal))^2)

R2_1 <- 1 - (SSR1)/SST1
R2_1
```

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

```
# summary(fit1)$r.squared
```

(vi)

β_0 measures the conditional mean of the log of an individual's annual salary in this bank for one who does not belong to an ethnic minority. This value is 10.396. Hence, this means that the conditional mean of an individual's annual salary for one who does not belong to an ethnic minority is $e^{10.396}$.

(vii)

```
LB1 <- coef(summary(fit1))[1, "Estimate"] -
  coef(summary(fit1))[1, "Std. Error"] * qt(1 - alpha_1/2, 472)

UP1 <- coef(summary(fit1))[1, "Estimate"] +
  coef(summary(fit1))[1, "Std. Error"] * qt(1 - alpha_1/2, 472)

confint1 <- cbind(LB1, UP1) %>%
  as.tibble() %>%
  rename("2.5 %" = LB1,
         "97.5 %" = UP1)

confint1 %>% kable()
```

2.5 %	97.5 %
10.35649	10.4363

```
#confint(fit1, level = 0.95)
```

(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.

(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.

1.d.

(i)

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

```
# OLS regression
fit2 <- lm(logsal ~ educ + logssal + gender + race + jobcat, data = dat)
fit2 %>% summary()
```

```
##
## Call:
## lm(formula = logsal ~ educ + logssal + gender + race + jobcat,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41156 -0.11677 -0.01430  0.09835  0.91176
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.025756   0.390102  10.320  < 2e-16 ***
## educ         0.024382   0.003653   6.675 7.01e-11 ***
## logssal      0.600624   0.044518  13.492  < 2e-16 ***
## gender       0.060849   0.018852   3.228  0.00134 **
## race        -0.042609   0.019187  -2.221  0.02685 *
## jobcat       0.120894   0.015717   7.692 8.65e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1666 on 468 degrees of freedom
## Multiple R-squared:  0.8261, Adjusted R-squared:  0.8242
## F-statistic: 444.6 on 5 and 468 DF,  p-value: < 2.2e-16
```

(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$_{race} = 0.027 > 0.001$

(iii)

```
# level of significance
alpha_2 <- 0.05

# unrestricted model
fit2.UR <- lm(logsal ~ educ + logssal + gender + race + jobcat, data = dat)

# restricted model
fit2.R <- lm(logsal ~ logsal, data = dat)

# Sum of squared residuals
SSR2.UR <- sum(resid(fit2.UR)^2)
SSR2.R <- sum(resid(fit2.R)^2)
```

```
# degrees of freedom
DF2_num <- df.residual(fit2.R) - df.residual(fit2.UR) #no. of restrictions = 5
DF2_denom <- df.residual(fit2.UR) #degrees of freedom of UR linear model = 468

# test statistic
tstat2 <- ((SSR2.R - SSR2.UR) / DF2_num) / ((SSR2.UR) / DF2_denom)
# ~ F(5, 468)

# critical value
fcrit2 <- qf(1 - alpha_2/2, DF2_num, DF2_denom)

# decision
tstat2 > fcrit2
```

```
## [1] TRUE
```

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* (conditional) 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.

(v)

```
coef(summary(fit2))[4, "Estimate"] - coef(summary(fit2))[5, "Estimate"]
```

```
## [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.

(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, this suggests that our new model has not overspecified and that this R^2 is not high largely because of adding more variables.

```
cbind(
  tibble("Model" = c("Old","New")),
         tibble("R-Squared" =
                  c(summary(fit1)$r.squared, summary(fit2)$r.squared))
) %>% kable()
```

Model	R-Squared
Old	0.0354137
New	0.8261007

```
vif(fit2) %>% as.tibble() %>% rename('VIF' = value) %>% kable()
```

VIF
1.892869
4.205978
1.505895
1.077052
2.517574

1.e.

We are working with the same model prior:

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

```
# level of significance
alpha_3 <- 0.1

# unrestricted model
fit3.UR <- lm(logsal ~ educ + logssal + gender + race + jobcat, data = dat)

# restricted model
fit3.R <- lm(logsal ~ educ + logssal + jobcat, data = dat)

# Sum of squared residuals
SSR3.UR <- sum(resid(fit3.UR)^2)
SSR3.R <- sum(resid(fit3.R)^2)
```

```r
# degrees of freedom
DF3_num <- df.residual(fit3.R) - df.residual(fit3.UR) #no. of restrictions = 2
DF3_denom <- df.residual(fit3.UR) #degrees of freedom of UR linear model = 468

# test statistic
tstat3 <- ((SSR3.R - SSR3.UR) / DF3_num) / ((SSR3.UR) / DF3_denom)
# ~ F(2, 468)

# critical value
fcrit3 <- qf(1 - alpha_3/2, DF3_num, DF3_denom)

# decision
tstat3 > fcrit3
```

```
## [1] TRUE
```

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.

1.d.

(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:

$$\hat{\beta_0} + \hat{\beta_1}\ educ + \hat{\beta_2}\ logssal + \hat{\beta_3}\ gender + \hat{\beta_4}\ race + \hat{\beta_5}\ jobcat$$

so, 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:

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

(ii)

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 who received the same starting salary as the individuals in population A, the average `logsal` is:

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

(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.

(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 $\beta_1 + \beta_4 = \beta_3$ 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)

```
#fit4.UR <- lm(logsal ~ educ + logssal + gender + race + jobcat, data = dat)

U <- dat$educ - dat$logssal
V <- dat$race - dat$educ

fit4.R <- lm(logsal ~ U + gender + V + jobcat, data = dat)

fit4.R %>% summary()
```

```
##
## Call:
## lm(formula = logsal ~ U + gender + V + jobcat, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46457 -0.12666 -0.01398  0.12058  0.85875
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.28172    0.16386  50.541  < 2e-16 ***
## U           -0.12024    0.01994  -6.029 3.34e-09 ***
## gender       0.16001    0.01916   8.353 7.57e-16 ***
## V           -0.16023    0.01860  -8.615  < 2e-16 ***
## jobcat       0.22745    0.01459  15.595  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1893 on 469 degrees of freedom
## Multiple R-squared:  0.7749, Adjusted R-squared:  0.773
## F-statistic: 403.6 on 4 and 469 DF,  p-value: < 2.2e-16
```

So, our restricted model is:

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

```
# level of significance
alpha_4 <- 0.1

# unrestricted model
fit4.UR <- lm(logsal ~ educ + logssal + gender + race + jobcat, data = dat)

# Sum of squared residuals
SSR4.UR <- sum(resid(fit4.UR)^2)
SSR4.R <- sum(resid(fit4.R)^2)

# degrees of freedom
DF4_num <- df.residual(fit4.R) - df.residual(fit4.UR) #no. of restrictions = 1
DF4_denom <- df.residual(fit4.UR) #degrees of freedom of UR linear model = 468

# test statistic
tstat4 <- ((SSR4.R - SSR4.UR) / DF4_num) / ((SSR4.UR) / DF4_denom)
# ~ F(1, 468)

# critical value
fcrit4 <- qf(1 - alpha_4/2, DF4_num, DF4_denom)

# decision
tstat4 > fcrit4
```

```
## [1] TRUE
```

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 a member of a ethnic minority and also has a starting salary.

1.g

(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 β_4.

(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 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.

1.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,

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

(ii)

The racial pay grap for males, controlling for education, starting salary, and job cateogry 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 $\beta_4 + \beta_6$.

(iii)

The racial pay grap for females, controlling for education, starting salary, and job cateogry 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 β_4.

(v)

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 ii. and part iii. this statement is equivalent to the stating that $\beta_4 + \beta 6 = \beta_4$ which in other words is to say that $\beta_6 = 0$,

```
# level of significance
alpha_5 <- 0.1

# OLS regression
fit5 <- lm(logsal ~ educ + logssal + gender + race + jobcat + race*gender,
           data = dat)

# test statistic
tstat5 <- coef(summary(fit5))[7, "Estimate"] /
  coef(summary(fit5))[7, "Std. Error"]
# ~t(467)

# critical value under the null
tcrit5 <- qt(1 - alpha_5/2, 467)

# decision
abs(tstat5) > tcrit5
```

```
## [1] FALSE
```

In conclusion, at 0.10 level of significance, we reject the null hypothes 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.