

HSCI 50 LAB 7: Correlation and Regression

INSTRUCTIONS

The data set that we will be using for the entire course is resampled data from the MIT COVID-19 Beliefs, Behaviors & Norms Survey (<https://covidsurvey.mit.edu/api.html>). This is a multi-country, online survey that examined different COVID-19 perceptions across time, from July 6, 2020 to March 28, 2021. We will be using data from the Philippines aged 20 to 60. **Assume the data are randomly sampled.**

In this lab, you will learn how to interpret correlation outputs between two continuous variables, and interpret outputs from simple and multiple linear and logistic regression models.

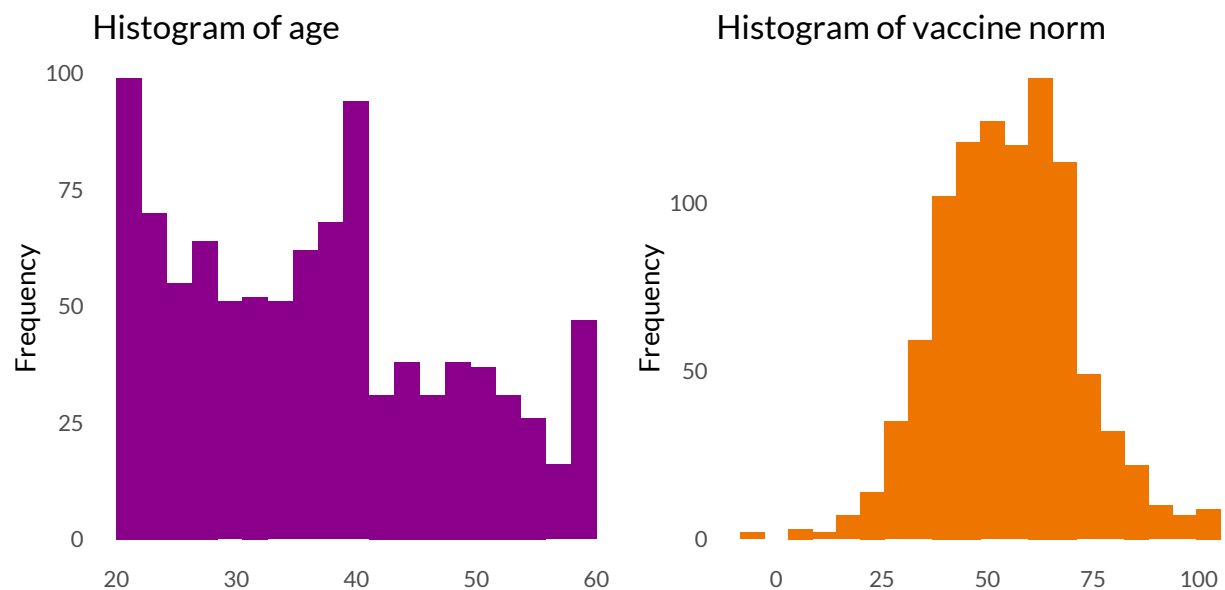
You have many options to submit this worksheet. Either you work on this by hand and scan/take a clear photo of your submission and save as PDF, or type your responses in a Word processor or PowerPoint presentation. You do not have to copy the questions again, but please number them accordingly.

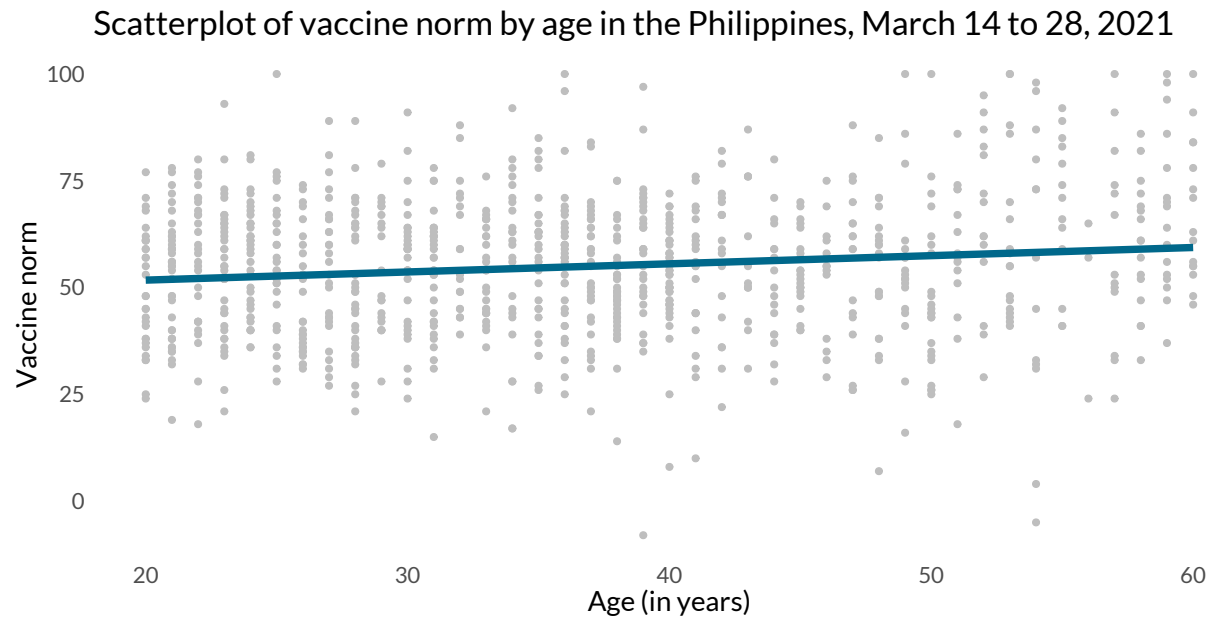
PART A. Correlation

Let us revisit the **vaccine norms** variable we had used in Labs 3, 4 and 6. Individuals were asked about their perceptions of how other people will accept the COVID-19 vaccine, as a measurement of vaccine norms. They were asked, "Out of 100 people in your community, how many do you think would take a COVID-19 vaccine if it were made available?" Again, while this is technically a discrete variable, we will assume for the purposes of this class that this is a continuous variable.

In Lab 6, we determined that vaccine norms differed by age group. We will examine the same relationship but with a different functional form of the age variable. We will use age as a continuous variable in years. While this is technically a discrete variable, we will assume for the purposes of this class that this is continuous.

The following are statistical outputs that will help you answer the following questions.





```
# Pearson's correlation coefficient output for vaccine norm by age relationship  
cor.test(x = data_norms_19$age, y = data_norms_19$vaccine_norm, method = 'pearson')
```

```
##  
## Pearson's product-moment correlation  
##  
## data: data_norms_19$age and data_norms_19$vaccine_norm  
## t = 4.0413, df = 959, p-value = 5.742e-05  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.06670931 0.19107788  
## sample estimates:  
## cor  
## 0.1294025
```

```
# Spearman's correlation coefficient output for vaccine norm by age relationship  
cor.test(x = data_norms_19$age, y = data_norms_19$vaccine_norm, method = 'spearman')
```

```
##  
## Spearman's rank correlation rho  
##  
## data: data_norms_19$age and data_norms_19$vaccine_norm  
## S = 133265743, p-value = 0.002111  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho  
## 0.09905126
```

1. What are the appropriate null and alternative hypotheses for the correlation above?
2. What is the scatterplot telling us? Is there some indication of a potentially linear relationship?
3. Based on the scatterplot, is it appropriate to calculate a Pearson's correlation coefficient, a Spearman's correlation coefficient, both, or neither?
4. What are the assumptions required for calculating the correlation coefficient?
5. What can you conclude? Either use the p -value or the 95% confidence interval (CI). Report the magnitude and direction of the correlation coefficient. Report Pearson's/Spearman's correlation coefficients as appropriate.

PART B. Linear Regression

We will now try to model the relationship of vaccine norm against age and gender. Gender is a binary variable in our dataset (Male / Female) while we will model age as a continuous variable centered at 20 years old, which is the lowest age in our data. Note that centering is not restricted to the mean of the data, we can center the value of any continuous variable to any constant value as appropriate to make the model more interpretable.

The following are statistical outputs that will help you answer the following questions. Aside from these statistical outputs, consider the statistical outputs from the previous question. Recall that centering a variable does not change the shape of the scatterplot, so the same scatterplot of vaccine norm by age applies to this part regardless of where we centered the variable.

When reading the statistical outputs of the linear regression model, pay attention to the Analysis of Variance table and the table of Coefficients. Under the table of Coefficients:

- (Intercept) shows the value of b_0
- The name of the variable shows the value of b_1
- The t -value is the t -statistic
- $\Pr(>|t|)$ shows the two-sided p -value of the t -statistic for the coefficient
- There are also 95% confidence intervals of the beta coefficients indicated under 2.5% for the lower limit of the confidence interval and 97.5% for the upper limit of the confidence interval.

The coefficient of determination is indicated under Multiple R-squared

There are also a set of diagnostic plots that should look familiar to you: a residuals vs. fitted values scatterplot, and a Q-Q plot of the residuals.

1. First, take a look at the statistical output of the simple linear regression of vaccine norm by age (Statistical Output 1). What are the appropriate null and alternative hypotheses for that particular linear regression model?
2. What are the assumptions required for the simple linear regression of vaccine norm by age (Statistical Output 1)? Do we expect the same set of assumptions for the simple linear regression of vaccine norm by gender?
3. Describe the equation of the line for all three regression models above, with the values of the coefficients from the regression model results. Recall that the format of the equation of the line for a simple linear regression is $Y = \beta_0 + \beta_1 X_1$ and add more $\beta_p X_p$ as needed in a multiple linear regression model.
4. For each of the three regression models above, what do they altogether say about the association of vaccine norm by age and gender? Either use the appropriate test statistic or the 95% confidence interval (CI). Report an interpretation for the point estimate of age in the simple linear regression model. Report an interpretation for the point estimate of age in the multiple linear regression as well.
5. Using the appropriate equation of the line from 3, predict the vaccine norm score for a 50 year old male. Report the point estimate only.

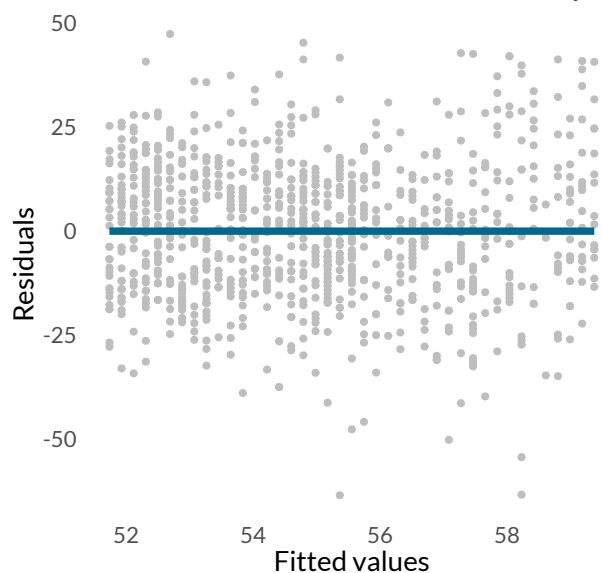
Statistical Outputs 1: Simple Linear Regression of Vaccine Norms by Age (centered at 20)

```
## Analysis of Variance Table
##
## Response: vaccine_norm
##           Df Sum Sq Mean Sq F value    Pr(>F)
## age_cent20  1  4348  4348.2   16.332 5.742e-05 ***
## Residuals 959 255321   266.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

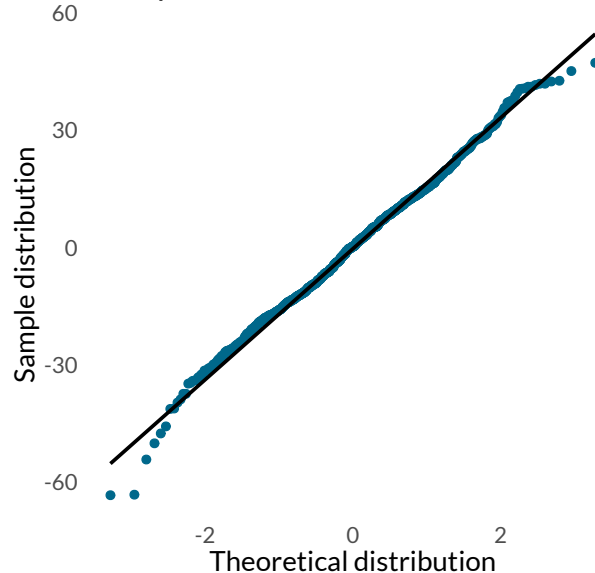
##
## Call:
## lm(formula = vaccine_norm ~ age_cent20, data = data_norms_19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -63.36 -11.60   0.30  11.02  47.31
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 51.73602    0.94256  54.889  < 2e-16 ***
## age_cent20   0.19093    0.04724   4.041 5.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.32 on 959 degrees of freedom
## Multiple R-squared:  0.01674,    Adjusted R-squared:  0.01572
## F-statistic: 16.33 on 1 and 959 DF,  p-value: 5.742e-05

##              2.5 %      97.5 %
## (Intercept) 49.88629391 53.5857464
## age_cent20   0.09821274  0.2836401
```

Residuals vs. Fitted values scatterplot



Q-Q plot of residuals



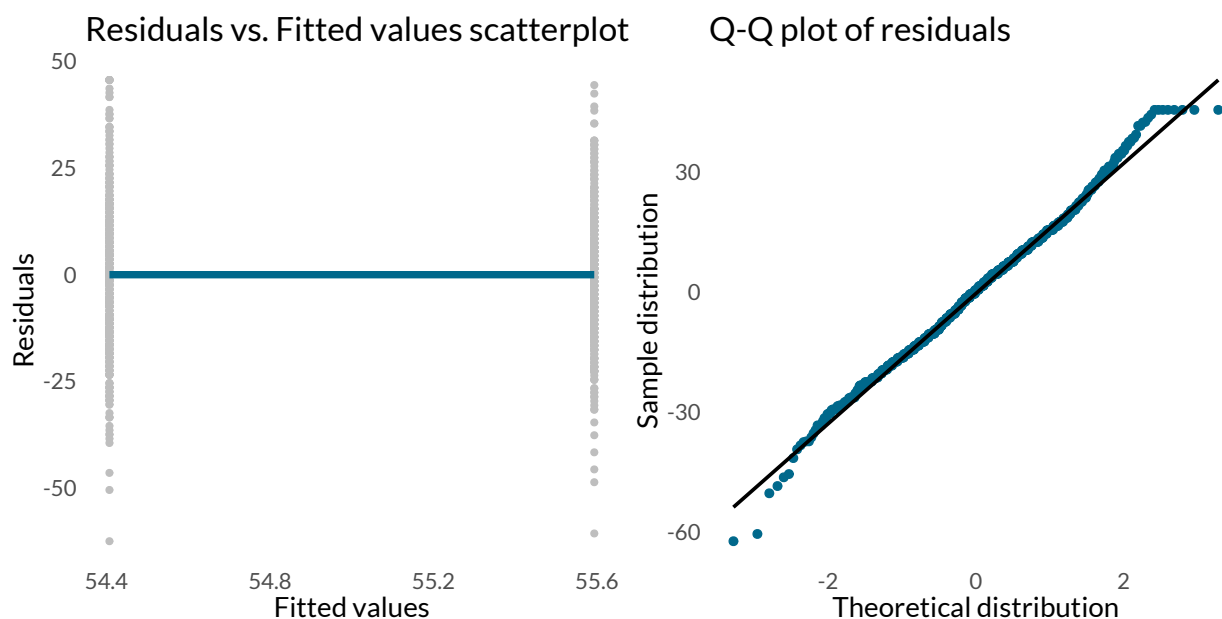
Statistical Outputs 2: Simple Linear Regression of Vaccine Norms by Gender

For gender, the reference category is female.

```
## Analysis of Variance Table
##
## Response: vaccine_norm
##           Df Sum Sq Mean Sq F value Pr(>F)
## gender      1    328   327.68   1.2117 0.2713
## Residuals 959 259342   270.43

##
## Call:
## lm(formula = vaccine_norm ~ gender, data = data_norms_19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -62.407 -11.407   0.407  10.593  45.593
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.4071     0.6918   78.642  <2e-16 ***
## gendermale    1.1864     1.0777    1.101    0.271
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.44 on 959 degrees of freedom
## Multiple R-squared:  0.001262, Adjusted R-squared:  0.0002205
## F-statistic: 1.212 on 1 and 959 DF, p-value: 0.2713

##              2.5 %   97.5 %
## (Intercept) 53.0493935 55.76477
## gendermale  -0.9286605  3.30137
```



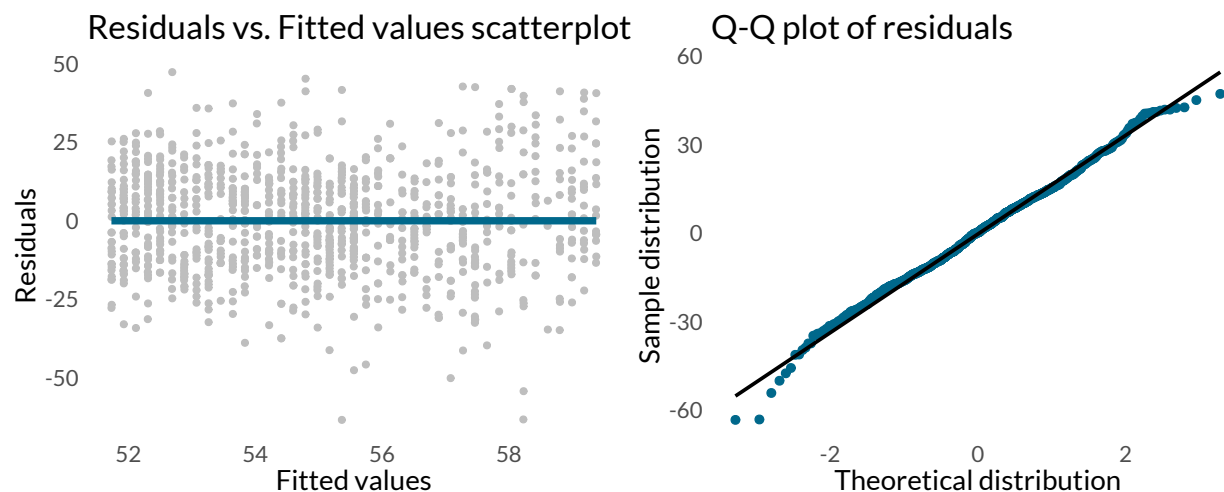
Statistical Outputs 3: Multiple Linear Regression of Vaccine Norms by Gender and Age (centered at 20)

For gender, the reference category is female.

```
## Analysis of Variance Table
##
## Response: vaccine_norm
##           Df Sum Sq Mean Sq F value    Pr(>F)
## gender      1     328    327.7   1.2295 0.2677831
## age_cent20   1    4020   4020.5  15.0854 0.0001098 ***
## Residuals  958  255321    266.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
## Call:
## lm(formula = vaccine_norm ~ gender + age_cent20, data = data_norms_19)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -63.365 -11.598   0.302  11.020  47.308
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.736833   0.971784  53.239  < 2e-16 ***
## gendermale   -0.003857   1.112938  -0.003  0.99724
## age_cent20    0.190973   0.049169   3.884  0.00011 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.33 on 958 degrees of freedom
## Multiple R-squared:  0.01675,    Adjusted R-squared:  0.01469
## F-statistic: 8.157 on 2 and 958 DF,  p-value: 0.000307

##              2.5 %      97.5 %
## (Intercept) 49.82976178 53.6439043
## gendermale  -2.18793526  2.1802210
## age_cent20   0.09448136  0.2874653
```



PART C. Logistic regression

For this part, we will recode the vaccine acceptance variable into a binary category similar to Lab 3 and 5 called vaccine hesitancy, where those who do not accept the vaccine or are unsure about it will be considered vaccine hesitant, otherwise they are not vaccine hesitant. Age is again analyzed as a continuous variable centered at age 20.

The following are statistical outputs that will help you answer the following questions.

When reading the statistical outputs of the logistic regression model, pay attention to the following:

- The `Coefficients` table is like the `Coefficients` table in the linear regression model result; however, the coefficients have not yet been exponentiated to odds ratios.
- The odds ratios and corresponding 95% CI are found at the end of the output under the table with the header columns 2.5% and 97.5%

Statistical Output 4: Simple Logistic Regression of Vaccine Hesitancy by Age (centered at 20)

```
##
## Call:
## glm(formula = hesitant ~ age_cent20, family = binomial(link = "logit"),
##      data = data_accept_19)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.3154  -1.1217  -0.8787   1.1738   1.5345
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.318697   0.104439   3.052  0.00228 **
## age_cent20  -0.028192   0.005341  -5.278  1.3e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1581.9  on 1144  degrees of freedom
## Residual deviance: 1553.2  on 1143  degrees of freedom
## AIC: 1557.2
##
## Number of Fisher Scoring iterations: 4

##              2.5 %    97.5 %
## (Intercept) 1.3753350 1.1213466 1.6890048
## age_cent20  0.9722019 0.9620009 0.9823662
```

Statistical Output 5: Simple Logistic Regression of Vaccine Hesitancy by Gender

For gender, the reference category is female.

```
##
## Call:
## glm(formula = hesitant ~ gender, family = binomial(link = "logit"),
##      data = data_accept_19)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.2805  -0.9344  -0.9344   1.0777   1.4417
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.23907    0.08077   2.960  0.00308 **
## gendermale  -0.84176    0.12201  -6.899 5.24e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1581.9  on 1144  degrees of freedom
## Residual deviance: 1533.1  on 1143  degrees of freedom
## AIC: 1537.1
##
## Number of Fisher Scoring iterations: 4

##              2.5 %   97.5 %
## (Intercept) 1.2700730 1.0845768 1.488777
## gendermale  0.4309495 0.3388883 0.546815
```

Statistical Output 6: Multiple Logistic Regression of Vaccine Hesitancy by Gender and Age (centered at 20)

For gender, the reference category is female.

```
##
## Call:
## glm(formula = hesitant ~ gender + age_cent20, family = binomial(link = "logit"),
##      data = data_accept_19)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.4451  -1.0652  -0.7725   1.1172   1.6904
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.610277   0.116311   5.247 1.55e-07 ***
## gendermale   -0.781868   0.123525  -6.330 2.46e-10 ***
## age_cent20   -0.024582   0.005459  -4.503 6.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1581.9  on 1144  degrees of freedom
## Residual deviance: 1512.4  on 1142  degrees of freedom
## AIC: 1518.4
##
## Number of Fisher Scoring iterations: 4

##              2.5 %    97.5 %
## (Intercept) 1.8409416 1.4676768 2.3161076
## gendermale   0.4575506 0.3587967 0.5823882
## age_cent20   0.9757179 0.9652671 0.9861572
```

For each of the three logistic regression models above, what do they altogether say about the association of vaccine hesitancy by age and gender? Either use the appropriate test statistic or the 95% confidence interval (CI). Provide an interpretation of the odds ratio between male and female from the simple logistic regression model and the multiple logistic regression model.

END OF LAB