Hannah LeBlanc (hkl362) Project 2 Revisions for **PROJECT 3**

Revisions are comments made in *red below the screenshots. I did not receive comments because I didn't submit my project 2 report through Gradescope and could not see comments on Canvas, so I just went based off of things I believed I could've done better based on the guidelines outlined on the rubric.

PART I: INTRODUCTION

```
title: "Project 2"
author: "Hannah LeBlanc"
date: "4/18/2021"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

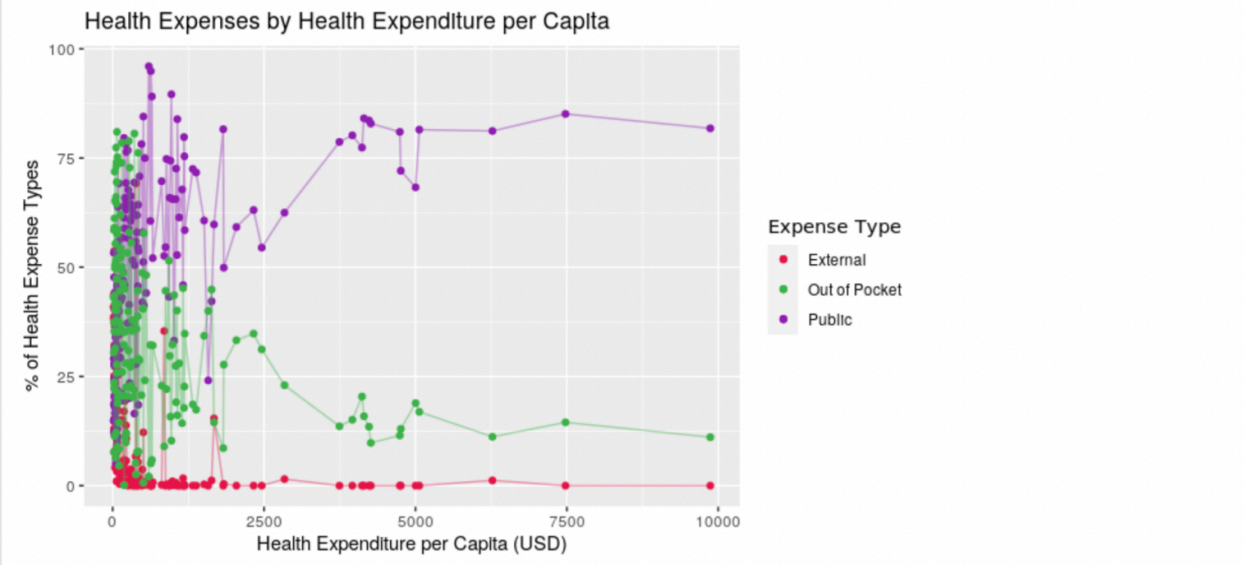
R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

Part 1: Introduction

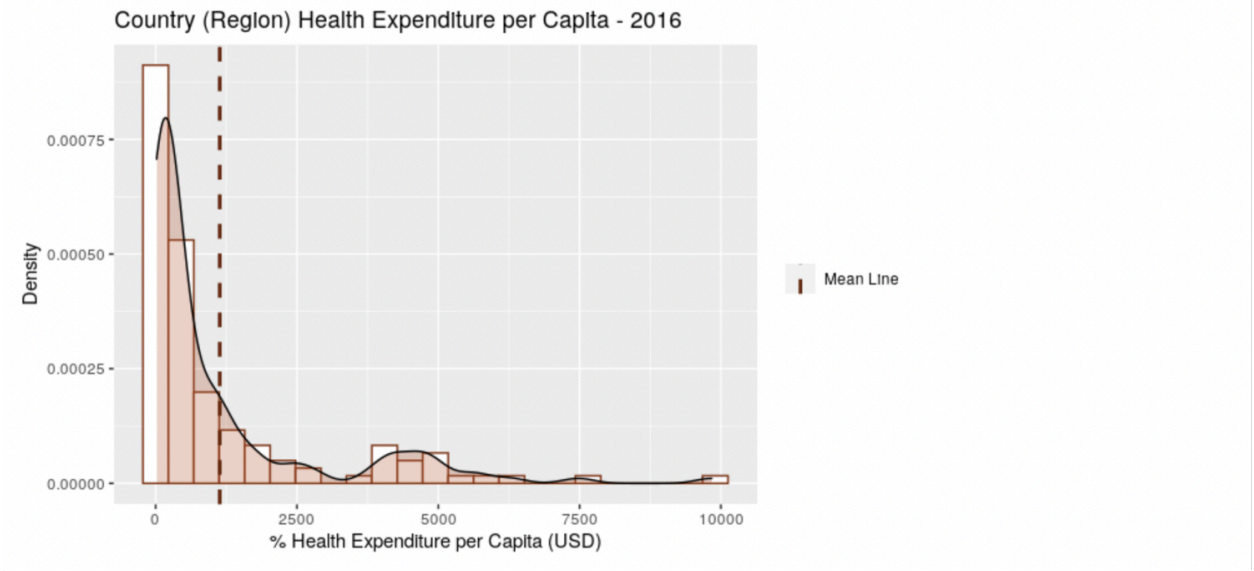
```{r, echo=FALSE, results=FALSE, message=FALSE}
#install.packages("readxl")
health_systems <- read_excel("health_systems.xlsx")
#view the dataset being used in the project
View(health_systems)
#create subset of dataset in which columns "Province_State" and "World_Bank_Name" are deleted
health_systems_p2 <- (subset(health_systems, select=-c(Province_State, World_Bank_Name)))
#view dataset with deleted columns
View(health_systems_p2)
```
The dataset I am using in my project is known as health_systems. The dataset includes variables including health expeditures per capita in countries around the globe, and the number of health care professionals (physicians, nurses, midwives, specialists) per capita. The dataset was originally described by World Bank- World Development Indicators: Health Systems. The World Bank gathers data by compiling international datasets, typically based on data generated by national statistical systems. For simplicity of completing the tasks outlined in this project, I deleted two columns, both categorical variables, leaving a single categorical variable out of the remaining 12 variables after removing two. I did not have to tidy the data any further because all rows represented observations. There were 210 total observations. This particular dataset interests me partially for the reason that it was compiled; the original statistician that collected this data was interested in seeing the affects of health spending and hospital staff on the spread of Covid-19. I am also particularly interested in the amount individuals are having to spend out of pocket in comparison to total health expenditures as well as the access they have to health care professionals. I would hope that individuals are not bearing the brunt of their health care costs, but fear I will observe the opposite. At the same time, I hope to observe that individuals who are having to pay out-of-pocket for their care have access to a range of health care providers.
```

*I felt like my introduction gave clear insight on my interest in the datasets I chose as well as the information I was hoping to test and better understand in my project. I gave a title, described the variables and provided a clear explanation as to why I was not including one column of data (part of my tidying step). I prepared the person who was viewing my project for what to expect in my report and what potential associations I was hoping to observe.

PART 2: EDA

```
-----------------------------------------------------------------------
### Part 2: EDA

#### Triple Line Graph
```{r}
health_expenditure_data <- health_systems_p2 %>%
  mutate_all(type.convert) %>%
  filter(!is.nan(Health_exp_public_pct_2016)) %>%
  filter(!is.nan(Health_exp_out_of_pocket_pct_2016)) %>%
  filter(!is.nan(External_health_exp_pct_2016)) %>%
  group_by(Health_exp_per_capita_USD_2016) %>%
    summarise(
      external_exp = min(External_health_exp_pct_2016),
      public_exp = min(Health_exp_public_pct_2016),
      outofpocket_exp = min(Health_exp_out_of_pocket_pct_2016)
    )

# Plot triple line graph
line_alpha <- 0.4
ggplot(health_expenditure_data, aes(x=Health_exp_per_capita_USD_2016)) +
  geom_line(aes(y = external_exp), color = "#e6194B", alpha = line_alpha) +
  geom_point(aes(y = external_exp, color = "External")) +
  geom_line(aes(y = public_exp), color = "#911eb4", alpha = line_alpha) +
  geom_point(aes(y = public_exp, color = "Public", labels="Public")) +
  geom_line(aes(y = outofpocket_exp), color = "#3cb44b", alpha = line_alpha) +
  geom_point(aes(y = outofpocket_exp, color = "Out of Pocket", labels="Out of Pocket")) +
  labs(title="Health Expenses by Health Expenditure per Capita",x="Health Expenditure per Capita (USD)", y = "% of Health Expense
ypes", colour="Expense Type") +
 theme(legend.title=element_text("Legend"), legend.box  ="vertical", legend.position = "right") +
 scale_color_manual(values = c("#e6194B","#3cb44b","#911eb4"))
```



*The triple line graph represents the three expense types-external, public and out-of-pocket- versus the health expenditure per capita of the countries included in the datasets. From the expenditure region on the x axis spanning from 0 to about 1500, there a sporadic array of expense values. Past this 1500 mark, you begin to see a pattern emerge. The public expense in countries with higher health expenditures per capita is the highest of the expense types. The out-of-pocket expenses seems lower than the pattern that appears in the 500 to 1500 range on the x axis. It is also very low compared to the public costs the higher the health expenditure per capita climbs. The external costs seem to stay close to zero no matter the change in the health expenditure per capita, with a few points deviating from this general pattern observed. *
```

```
#### Histogram
```{r}
# Histogram with density plot
histodata <- health_happiness %>%
  filter(Health_exp_per_capita_USD_2016 != NaN) %>%
  mutate(Health_exp_per_capita_USD_2016 = as.numeric(Health_exp_per_capita_USD_2016))

#hist(tidydf$Health_exp_per_capita_USD_2016)
ggplot(histodata, aes(x=Health_exp_per_capita_USD_2016)) +
  geom_histogram(aes(y=..density..), binwidth=450, color="#80381a", fill="white") +
  geom_density(alpha=0.25, fill="#b34e24") +
  geom_vline(aes(xintercept=mean(Health_exp_per_capita_USD_2016), color="Mean Line", label="Mean"), linetype="dashed", size=1) +
  labs(
    title="Country Health Expenditure per Capita - 2016",
    x="% Health Expenditure per Capita (USD)",
    y = "Density"
  ) +
  theme(legend.title = element_blank()) +
  scale_color_manual(values = c("#662d15"))
```
```



This histogram represents the distribution of % health expenditure per capita in the countries included in the datasets. It is right skewed with a mean value of about 1250.

---

*I created two visualizations of important variables involved in my project. The first of which was health expenditures by expense type. The second was a visualization of the % Health Expenditure per Capita by country/region. I was able to use these visualizations in order to display relationships between some of the variables in my dataset. However, I did not create summary statistics for any of my variables. In order to complete this section, I may add summary statistics including mean values, standard error, etc. for variables of interest like external expenses, out of pocket expenses, public expenses, and total health expenditure, which were all considered in the first figure I included for visualization.

PART 3: MANOVA

```{r, echo=FALSE, results=FALSE, message=FALSE}
# How do the means of level of current health expenditure expressed as a percentage of GDP  and share of out-of-pocket payments of total current health expenditures differ between countries?
health_systems_p2 %>%
  group_by(Country_Region) %>%
  summarize(Health_exp_pct_GDP_2016, Health_exp_out_of_pocket_pct_2016)

# Represent the means per Country
health_systems_p2 %>%
  select(Country_Region,Health_exp_pct_GDP_2016,Health_exp_out_of_pocket_pct_2016) %>%
  pivot_longer(-1,names_to='DV', values_to='measure') %>%
  ggplot(aes(Country_Region,measure,fill=Country_Region)) +
  geom_bar(stat="summary", fun = "mean") +
  geom_errorbar(stat="summary", fun.data = "mean_se", width=.5) +
  facet_wrap(~DV, nrow=2) +
  coord_flip() +
  ylab("") +
  theme(legend.position = "none")

# Inspect multivariate plots of response variable for each Country
ggplot(health_systems_p2, aes(x = Health_exp_pct_GDP_2016, y = Health_exp_out_of_pocket_pct_2016)) +
  geom_point(alpha = .5) +
  geom_density_2d(h=2) +
  coord_fixed() +
  facet_wrap(~Country_Region)

 # Inspect homogeneity of (co)variances
 covmats <- health_systems_p2 %>%
   group_by(Country_Region) %>%
   do(covs=cov(.[2:3]))

 # Covariance matrices per Country
 for(i in 1:3){print(as.character(covmats$Country_Region[i])); print(covmats$covs[i])}
```

```{r manova}
# Perform MANOVA with 2 response variables listed in cbind()
manova_hs2 <- manova(cbind(Health_exp_pct_GDP_2016, Health_exp_out_of_pocket_pct_2016) ~ Country_Region, data = health_systems_p2)

# OUtput of MANOVA
summary(manova_hs2)

  # If MANOVA is significant then we can perform one-way ANOVA for each variable
  summary.aov(manova_iris)

    # The MANOVA was not significant therefore I do not perform post-hoc analysis

    # Perform Bonferonni correction to adjust value of alpha
    #alpha prime = alpha/ number of tests
  0.05/2
```

```
[1] 0.025
```

I performed a MANOVA in hopes of determining if the means of the variables representing the level of current health expenditure expressed as a percentage of GDP and share of out-of-pocket payments of total current health expenditures differ between countries included in this study. My results were not signficant, therefore I did not perform univariate analysis or post hoc analysis. I performed the MANOVA and the summary.aov so the number of tests i used in order to determine the Bonferonni correction was 2. The adjusted significance value was 0.025.The overall type I error remains at 0.05 after this adjustment. The assumptions for the MANOVA include a random sample with independent observations, multivariate normality of numeric response variables, homogeneity of within-groups covariance matrices, linear relationships among variables but not collinearity and no extreme univariate or multivariate outliers. It is highly likely that the dataset includes random samples with independent observations as it is a major organization's published data that is shared with the public. The multivariate normality and within-group homogeneity assumptions are also met as coded for above. There are no extreme outliers.

*In this section I performed a MANOVA that according to my description was not significant. Because of this, I did not perform ANOVAs of post-hoc t tests. I did not have to correct my data so I did not interpret p values again. I discussed the assumptions that had to be met in order to perform such tests.

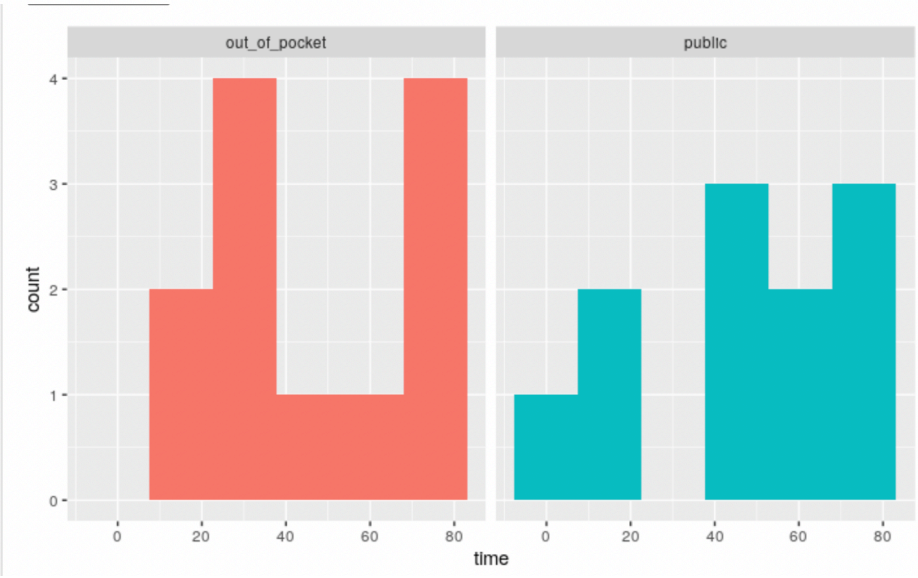
PART 4: RANDOMIZATION TEST

```{r manova}
# Example of expenditures from the health_systems_2 dataset
out_of_pocket <- c(77.4,58,30.9,41.7,35.2,15.8,80.60,18.90,78.90,27.7,28, 71.9)
public <- c(5.10, 41.4, 67.7, 49.1, 44.1, 60.06, 74.4, 16.5, 68.3, 72.5, 20)

# Save in a data frame, create the variables condition (out_of_pocket/public) and time
expenditures <- data.frame(condition = c(rep("out_of_pocket",12),
                                         rep("public",11)),
                           time = c(out_of_pocket,public))
head(expenditures)
```

	condition <fctr>	time <dbl>
1	out_of_pocket	77.4
2	out_of_pocket	58.0
3	out_of_pocket	30.9
4	out_of_pocket	41.7
5	out_of_pocket	35.2
6	out_of_pocket	15.8

6 rows

```{r explore}
# Represent the distribution of time for each condition
ggplot(expenditures, aes(time,fill=condition)) +
  geom_histogram(bins=6.5) +
  facet_wrap(~condition,ncol=2) +
  theme(legend.position="none")

# Calculate the mean difference between the two conditions
true_diff <- expenditures %>%
  group_by(condition) %>%
  summarize(means = mean(time)) %>%
  summarize(mean_diff = diff(means)) %>%
  pull
true_diff
```



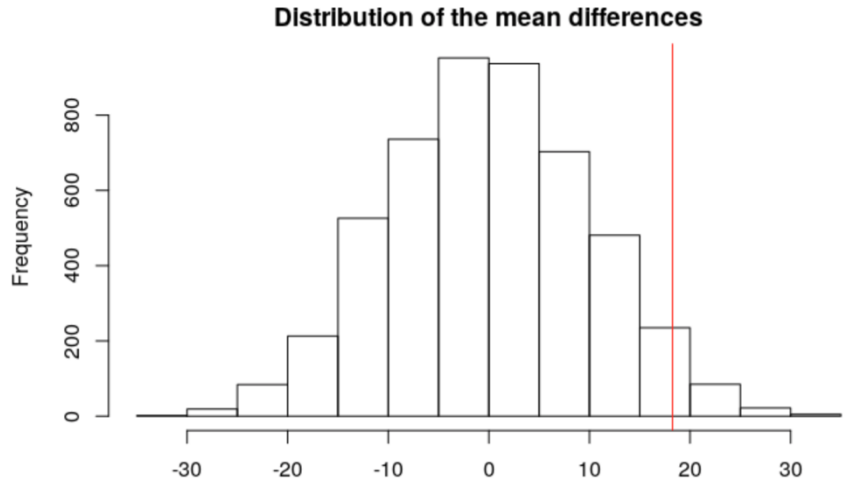
```{r randomize}
# Keep the same condition, resample the time across conditions
perm1 <- data.frame(condition = expenditures$condition, time = sample(expenditures$time))
head(perm1)

# Find the new mean difference
perm1 %>%
  group_by(condition) %>%
  summarize(means = mean(time)) %>%
  summarize(mean_diff = diff(means))

## Repeat randomization
# Keep the same condition, resample the time across conditions
perm2 <- data.frame(condition = expenditures$condition, time = sample(expenditures$time))
head(perm2)

# Find the new mean difference
perm2 %>%
  group_by(condition) %>%
  summarize(means = mean(time)) %>%
  summarize(mean_diff = diff(means))
```

```{r}
## Repeat randomization many times
# Create an empty vector to store the mean differences
mean_diff <- vector()

# Create many randomizations with a for loop
for(i in 1:5000){
  temp <- data.frame(condition = expenditures$condition, time = sample(expenditures$time))

  mean_diff[i] <- temp %>%
    group_by(condition) %>%
    summarize(means = mean(time)) %>%
    summarize(mean_diff = diff(means)) %>%
    pull
}
```

mean_diff <dbl>
0.5730303

1 row

```{r pvalue}
# Represent the distribution of the mean differences with a vertical line showing the true difference
{hist(mean_diff, main="Distribution of the mean differences"); abline(v = 18.258, col="red")}

# Calculate the corresponding two-sided p-value
mean(mean_diff > -true_diff | mean_diff < true_diff)
#mean(mean_diff > 18.258 | mean_diff < -18.258)
```




```{r ttest}
# Compare to a Welch's t-test
t.test(data = expenditures, time ~ condition)
```

```
        Welch Two Sample t-test

data:  time by condition
t = -0.011043, df = 20.886, p-value = 0.9913
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -21.40638  21.18032
sample estimates:
mean in group out_of_pocket        mean in group public
                   47.08333                    47.19636
```

I performed a randomization test on the out of pocket and public health expenditures. The null hypothesis was that the mean expenditures were the same for out of pocket versus public costs. The alternate hypothesis was that the mean expenditures for out of pocket versus pubilc costs were different. Based on the high p value of 0.9913, I fail to reject the null hypothesis and conclude that the mean expenditures out of pocket and public were the same.

---

*I successfully performed a randomization test and included the hypotheses in my description below the test. I included the distribution of mean differences and test statistic (p value) along with interpretations.

PART 5: LINEAR REGRESSION

```
---------------------------------------------------------------------
### Part 5: Linear Regression Model

```{r, echo=FALSE, results=FALSE, message=FALSE}
#call dataset into view
View(health_systems_p2$`Physicians_per_1000_2009-18`)

# mean center numeric variables involved
fixed_physicians <-na.omit(health_systems_p2$`Physicians_per_1000_2009-18`)
fixed_pocket <- na.omit(health_systems_p2$Health_exp_out_of_pocket_pct_2016)

physicians_c = mean(fixed_physicians, na.rm = TRUE)
pocket_c = mean(fixed_pocket, na.rm =  TRUE)

# Let's visualize the relationships
ggplot(health_systems_p2, aes(x = physicians_c, y = pocket_c)) +
  geom_point() +
  geom_smooth(method = lm, se = FALSE, fullrange=TRUE)
```
```

```
### b. Model

```{r model}
# Fit a model with physicians_c as reference
fit <- lm(pocket_c ~ physicians_c, data = health_systems_p2)
summary(fit)
```
```

```
Error in summary(fit) : object 'fit' not found
```

```
### c. Assumptions

```{r assumptions}
# Check assumptions
# Linearity and homoscedasticity
plot(fit, which = 1)
bptest(fit)

# Normality
plot(fit, which = 2)
shapiro.test(fit$residuals)
```

d. Bootstrap SEs

```{r bootstrapping}
##  Bootstrap from observations
# Repeat bootstrapping 5000 times, saving the coefficients each time
samp_SEs <- replicate(5000, {
  # Bootstrap your data (resample observations)
  boot_data <- sample_frac(med, replace = TRUE)
  # Fit regression model
  fitboot <- lm(physicians_c ~ pocket_c, data = boot_data)
  # Save the coefficients
  coef(fitboot)
})
```
```

```r
# Estimated SEs
samp_SEs %>%
  # Transpose the obtained matrices
  t %>%
  # Consider the matrix as a data frame
  as.data.frame %>%
  # Compute the standard error (standard deviation of the sampling distribution)
  summarize_all(sd)

##  Bootstrap from residuals
# Repeat bootstrapping 5000 times, saving the coefficients each time
resids_SEs <- replicate(5000, {
  # Bootstrap your residuals (resample with replacement)
  new_resids <- sample(fit$residuals, replace = TRUE)
  # Consider a new response as fitted values plus residuals
  boot_data <- med
  boot_data$new_y = fit$fitted.values + new_resids
  # Fit regression model
  fitboot <- lm(new_y ~ Age_c * BMI_c, data = boot_data)
  # Save the coefficients
  coef(fitboot)
})

# Estimated SEs
resids_SEs %>%
  # Transpose the obtained matrices
  t %>%
  # Consider the matrix as a data frame
  as.data.frame %>%
  # Compute the standard error (standard deviation of the sampling distribution)
  summarize_all(sd)
```

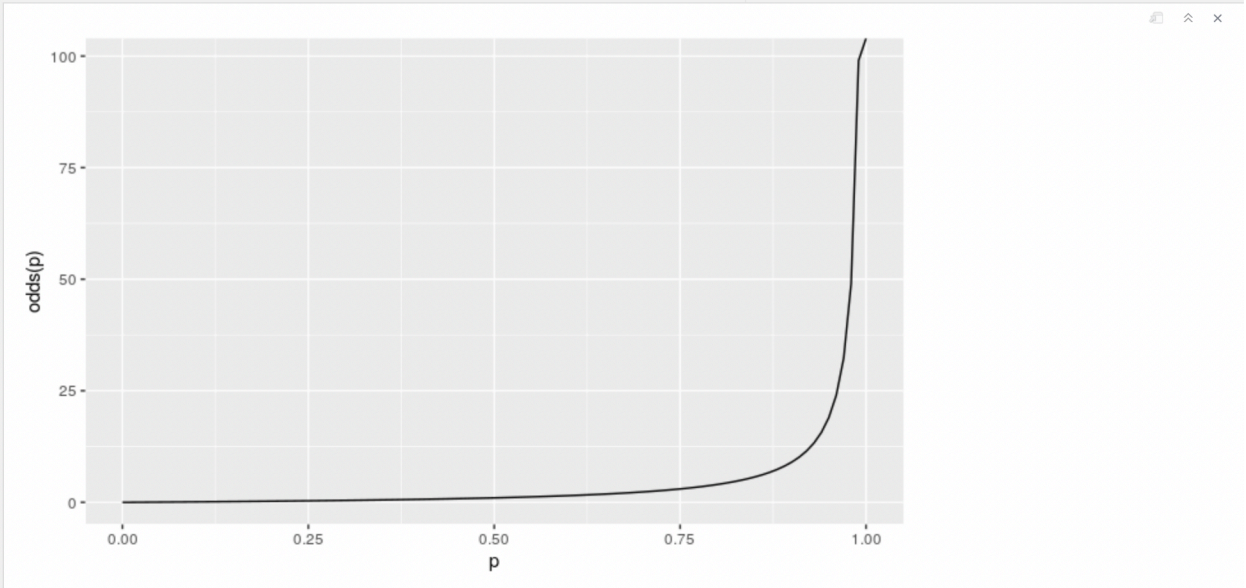
```{r comparison}
# Compare with normal-theory SEs
coeftest(fit)[,1:2]

# Compare with robust SEs
coeftest(fit, vcov = vcovHC(fit))[,1:2]

```

I included the code needed in order to carry out the linear regression and all requirements. I was unable to make the code work with my data and therefore do not have interpretation for interactions or variations. The assumptions for a linear regression include that it is a linear function (the mean response is a linear function of Xi), errors are independent, the errors are normally distributed and the errors have equal variances. In order to visually carry out assumption testing, a residual plot, histogram and QQ plot must be generateed. If the residual plot results in residuals that are randomly dispersed across the horizontal axis, the linear regression is a good fit for the data in question. A histogram is then constructed in order to determine whether or not the data is normally distributed. If the histogram reflects a normal curve, with no skew, it passes the assumption for a linear regression that errors are normally distributed. A Q-Q Plot is then constructed and if the residuals are linear, as in they fall along and near the linear line on the Q-Q Plot, assumptions have been met for the linear regression to be carried out. Both Shapiro-Wilk and Kolmogorov-Smirnov tests can be carried out the test normality using the null hypothesis that the distribution is normal. Equal variance or homoscedasticity can be tested using the Breusch-Pagan test in which the null hypothesis is that homoscedasticity exists. If all assumptions are met, a linear regression can be carried out. Robust standard errors are computed if the assumption of homoscedasticity fails to apply to the data. In my dataset I was to calculate them regardless of meeting assumptions. The goal of robust standard errors is to obtain ubiased error values. Based on my inability to carry out the linear regression and its associated tests, I am unsure as to whether or not my data violated any assumptions. However, if it had, I would turn to bootstrapping. Bootstrapping samples are used in order to estimate coefficients, SEs and other fitted values. Bootstrapped values for standard error can be compared to both the original and robust SEs calculated earlier.

PART 6: LOGISTIC MODEL

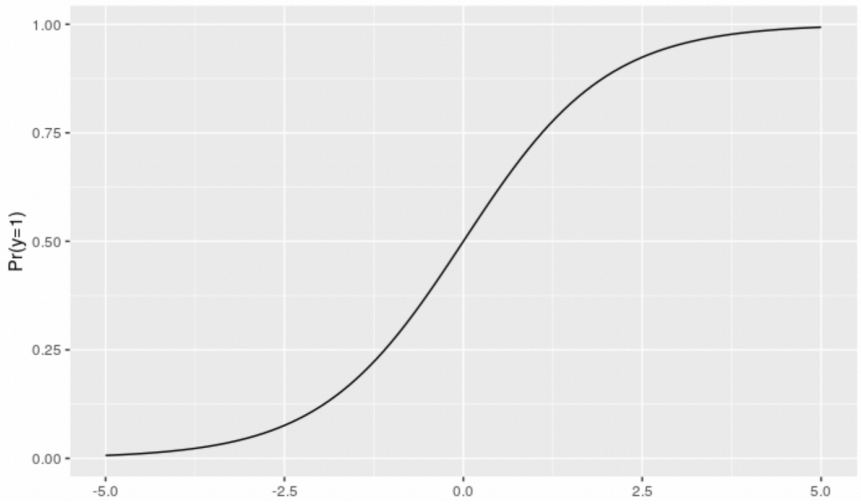
```
---------------------------------------------------------------------
### Part 6: Logistic Regression

### a. Odds
```

```{r odds}

# Simulate probability values (varying between 0 and 1 by 0.1)
p <-seq(0, 1, by = .1)

# Create a dataframe with these probabilities and corresponding odds
cbind(p, odds = odds(p)) %>%
  round(4) %>%
  as.data.frame %>%
  # Represent the relationship between probabilities and odds
  ggplot() +
  stat_function(aes(p), fun = odds, geom="line") +
  ylab("odds(p)") + xlab("p")
```




```
### b. Logit function
```

```{r logit}
# We also need to define the logit link function (logarithm of odds)
logit <- function(p) log(odds(p))
cbind(p, odds = odds(p), logit = logit(p)) %>%
  round(4) %>%
  as.data.frame %>%
  # Represent the relationship between probabilities and odds
  ggplot() +
  stat_function(aes(p), fun = logit, geom="line") +
  ylab("logit(p)") + xlab("p")

# Logistic model
logistic <- function(x){exp(x) / (1 + exp(x))}
x <- seq(-5,5, by = .1)
cbind(x, model = logistic(x)) %>%
  as.data.frame %>%
  ggplot() +
  stat_function(aes(x), fun = logistic, geom="line") +
  xlab("x") + ylab("Pr(y=1)")
```



c. Model

```{r model}
#create a binary variable coded as 0 and 1
expenditures_GDP <- health_systems_p2 %>%
  mutate(Country_Region=ifelse(,1,0))

# Fit a new regression model
fit <- glm(y ~ Health_exp_pct_GDP_2016, data = health_systems_p2, Country_Region = binomial(link="logit"))
summary(fit)

# Interpret the coefficients by considering the odds (inverse of log(odds))
exp(coef(fit))
```

```
Error: Problem with `mutate()` input `Country_Region`. x argument "test" is missing, with no default ℹ Input
 `Country_Region` is `ifelse(, 1, 0)`. Run `rlang::last_error()` to see where the error occurred.
```

Show Traceback

d. Fitted values

```{r logistic}
# Predicted logit (log-odds) for Health_exp_pct_GDP_2016 = 10
predict(fit, newdata = data.frame(Health_exp_pct_GDP_2016 = 10), type = "link")

# Predicted odds
exp(predict(fit, newdata = data.frame(Health_exp_pct_GDP_2016 = 10), type = "link"))

# Predicted probability for Health_exp_pct_GDP_2016 = 10
predict(fit, newdata = data.frame(Health_exp_pct_GDP_2016 = 10), type = "response")
```

```
Error in predict(fit, newdata = data.frame(Health_exp_pct_GDP_2016 = 10),  :
  object 'fit' not found
```

e. Represent model

```{r plotmodel}
# Add predicted probabilities to the dataset
health_systems_p2$Health_exp_pct_GDP_2016 <- predict(fit, type = "response")

# Predicted outcome is based on the probability of
health_systems_p2$Health_exp_pct_GDP_2016 <- ifelse(health_systems_p2$Health_exp_pct_GDP_2016 > .5)

# Plot the model
ggplot(health_systems_p2 aes(Health_exp_pct_GDP_2016,y)) +
  geom_jitter(aes(color = predicted), width = .3, height = 0) +
  stat_smooth(method="glm", method.args = list(family="binomial"), se = FALSE) +
  geom_hline(yintercept = 0.5, lty = 2) +
  ylab("Pr")

# Save the predicted log-odds in the dataset
health_systems_p2$logit <- predict(fit)

# Compare to the outcome in the dataset with a density plot
ggplot(health_systems_p2, aes(logit, fill = as.factor(outcome))) +
  geom_density(alpha = .3) +
  geom_vline(xintercept = 0, lty = 2)

```

My coding for the logistic model did not work. I tried my to best explain the model without having my data to use in the explanation. A logistic regression considers the probabilities and odds ratios, as I included the code for above. It uses both numeric and categorical predictors to model the probability of an outcome. In my dataset I used the Country_Region variable as my categorical variable and used the Health_exp_pct_GDP_2016 as my numeric variable. The binary response was determined and modeled the probability of the outcome occuring.I began by simulatin odds and create a dataframe with these probabilities and corresponding odds before creating a visual representation of the relationship between the odds and probabilities. I carried out a logit link function and further explored the relationship between proabilities and odds. I then created a binary variable using my Country_Region variable and making it into values of 0 and 1 only. With this, I create a regression model and, if I had numerical answers, I would interpret the meaning of the coefficients in context. I deleted the portion that involves analysis of the model as I have no values to enter in, however I would have calculated the accuracy, sensitivity, specificity and precision of the regression. Accuracy decribes the proportion of correctly classified groups. Sensitivity is the proportion of true positives and specificity is the proportion of true negatives. Precision is the proportion of true positive prediction within the model. At the end I would've plotted the model in order to show visually the results of the logistic regression.

---

*COMBINED PART 4 AND 5 REVISIONS: Out of the 125 points allotted to this project, I received 88 points. I lost 37 points from the two sections above. In part 4, I was unable to carry out a linear regression, however I described the assumptions and how I would go about testing them. Admittedly, I was unable to determine whether my regression passed assumptions and was not able to move forward with the test. In order to improve this section, I may try to interpret different variables. I thought it was a possibility that I did not choose two variables that had true correlations and therefore was unable to carry out the regression. With a better choice of variables, or tweaking to the code I provided, my linear regression may have been more successful. As I said in my description, my next step if I would've found that my data violated assumptions, I would have turned to bootstrapping. I did not calculate the robust or bootstrapped SEs. For Part 5, I did not build a logistic regression as my code again did not work. I did not construct an ROC curve, determine the AOC or interpret coefficients because I did not successfully perform the regression.

*It is also important to note that I most likely lost points on formatting, as I was supposed to turn in a knitted pdf and I turned in screenshots of my code.