Biostat 200C Homework 1

Due Apr 14 @ 11:59PM

Contents

To submit homework, please submit Rmd and html files to bruinlearn by the deadline.

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```r
library(datasets)
library(gtsummary)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from   
##   +.gg   ggplot2
```

```
library(leaps)
library(faraway)
```

```
## Warning in check_dep_version(): ABI version mismatch: 
## lme4 was built with Matrix ABI version 1
## Current Matrix ABI version is 0
## Please re-install lme4 from source or restore original 'Matrix' package
```

```
## 
## Attaching package: 'faraway'
## 
## The following object is masked from 'package:GGally':
## 
##     happy
```

Q1. Reivew of linear models

The swiss data — use Fertility as the response to practice

- An initial data analysis that explores the numerical and graphical characteristics of the data.

```
swiss |> head(10)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
## Broye             83.8        70.2          16         7    92.85
## Glane             92.4        67.8          14         8    97.16
## Gruyere           82.4        53.3          12         7    97.67
## Sarine            82.9        45.2          16        13    91.38
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
## Broye                    23.6
## Glane                    24.9
## Gruyere                  21.0
## Sarine                   24.4
```

```
swiss <- swiss |>
  as_tibble() |>
  print(width = Inf)
```

```
## # A tibble: 47 x 6
##    Fertility Agriculture Examination Education Catholic Infant.Mortality
##        <dbl>       <dbl>       <int>     <int>    <dbl>            <dbl>
##  1      80.2        17            15        12     9.96             22.2
##  2      83.1        45.1           6         9    84.8              22.2
##  3      92.5        39.7           5         5    93.4              20.2
##  4      85.8        36.5          12         7    33.8              20.3
##  5      76.9        43.5          17        15     5.16             20.6
##  6      76.1        35.3           9         7    90.6              26.6
##  7      83.8        70.2          16         7    92.8              23.6
##  8      92.4        67.8          14         8    97.2              24.9
##  9      82.4        53.3          12         7    97.7              21
## 10      82.9        45.2          16        13    91.4              24.4
## # i 37 more rows
```

```
str(swiss)
```

```
## tibble [47 x 6] (S3: tbl_df/tbl/data.frame)
##  $ Fertility       : num [1:47] 80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
##  $ Agriculture     : num [1:47] 17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
##  $ Examination     : int [1:47] 15 6 5 12 17 9 16 14 12 16 ...
##  $ Education       : int [1:47] 12 9 5 7 15 7 7 8 7 13 ...
##  $ Catholic        : num [1:47] 9.96 84.84 93.4 33.77 5.16 ...
##  $ Infant.Mortality: num [1:47] 22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

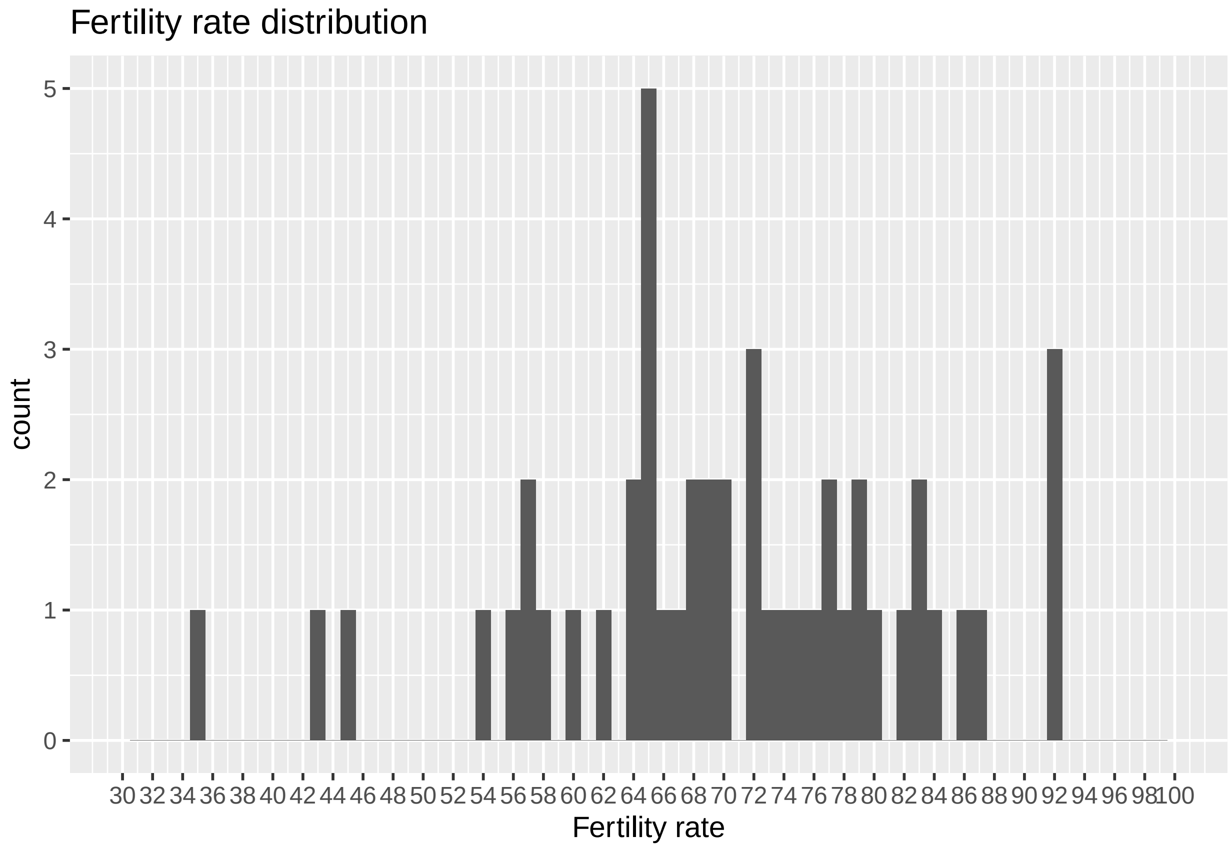
```
swiss |>
  tbl_summary() |>
  bold_labels()
```

```
## Table printed with ‘knitr::kable()‘, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include ‘message = FALSE‘ in code chunk header.
```

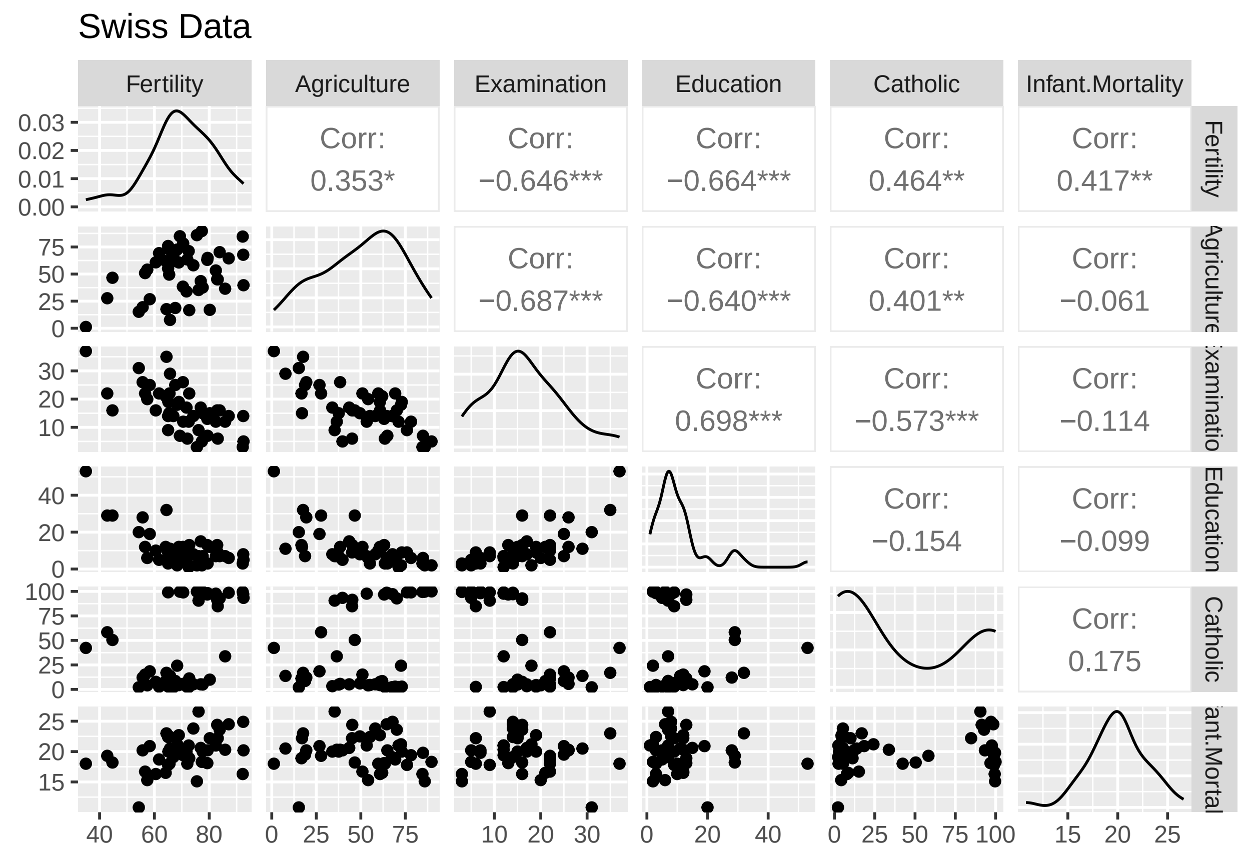
Characteristic	**N = 47**
Fertility	70 (65, 78)
Agriculture	54 (36, 68)
Examination	16 (12, 22)
Education	8 (6, 12)
Catholic	15 (5, 93)
Infant.Mortality	20.00 (18.15, 21.70)

```
ggplot(data = swiss) +
  geom_histogram(binwidth = 1,aes(x = Fertility)) +
  scale_x_continuous(breaks = seq(30, 100, 2), lim = c(30, 100)) +
  xlab('Fertility rate') +
  ggtitle("Fertility rate distribution")
```

```
## Warning: Removed 2 rows containing missing values (‘geom_bar()‘).
```



```
ggpairs(data = swiss) +
  labs(title = "Swiss Data")
```



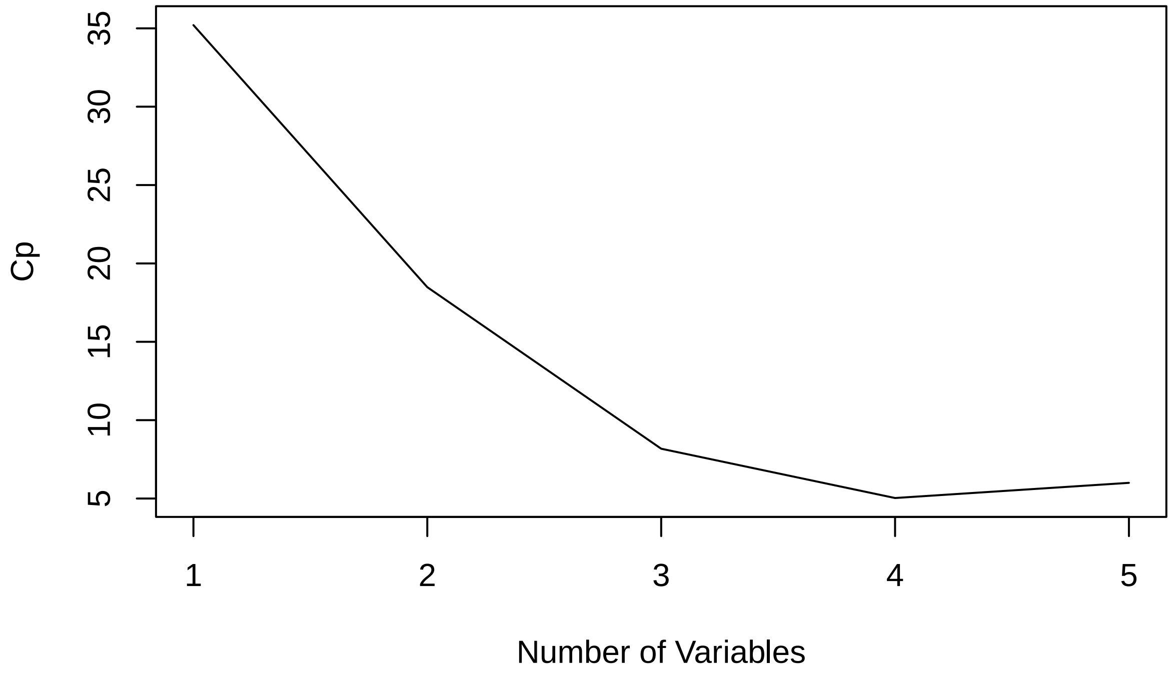
Answer: By initial data exploration, although there contains 2 missing values in fertility, the data is relatively clean. The range of each seems to make sense by observing the numerical analysis. For graphical analysis, the histogram looks ok. There is no obvious outliers exist. Since we only have 47 observations in total, it is also hard for us to conclude on the general trend of fertility based on the histogram. The paired scatter plot shows that fertility might positively associated with agriculture and infant mortality, negatively associated with examination, education. The relationship between fertility and catholic is not clear from the plot.

- Variable selection to choose the best model.

```
regfit_full <- regsubsets(Fertility ~ ., data = swiss)
reg_summary <- summary(regfit_full)
print(reg_summary)
```

```
## Subset selection object
## Call: regsubsets.formula(Fertility ~ ., data = swiss)
## 5 Variables  (and intercept)
##                  Forced in Forced out
## Agriculture          FALSE      FALSE
## Examination          FALSE      FALSE
## Education            FALSE      FALSE
## Catholic             FALSE      FALSE
## Infant.Mortality     FALSE      FALSE
## 1 subsets of each size up to 5
## Selection Algorithm: exhaustive
##           Agriculture Examination Education Catholic Infant.Mortality
## 1  ( 1 ) " "         " "         "*"       " "      " "
## 2  ( 1 ) " "         " "         "*"       "*"      " "
## 3  ( 1 ) " "         " "         "*"       "*"      "*"
## 4  ( 1 ) "*"         " "         "*"       "*"      "*"
## 5  ( 1 ) "*"         "*"         "*"       "*"      "*"
```

```
plot(reg_summary$cp, xlab = "Number of Variables", ylab = "Cp", type = "l")
```



Answer: 4 variables are selected by Cp criterion. It includes Agriculture, Education, Catholic, and Infant.Mortality. We will be using these 4 variables to fit the model.

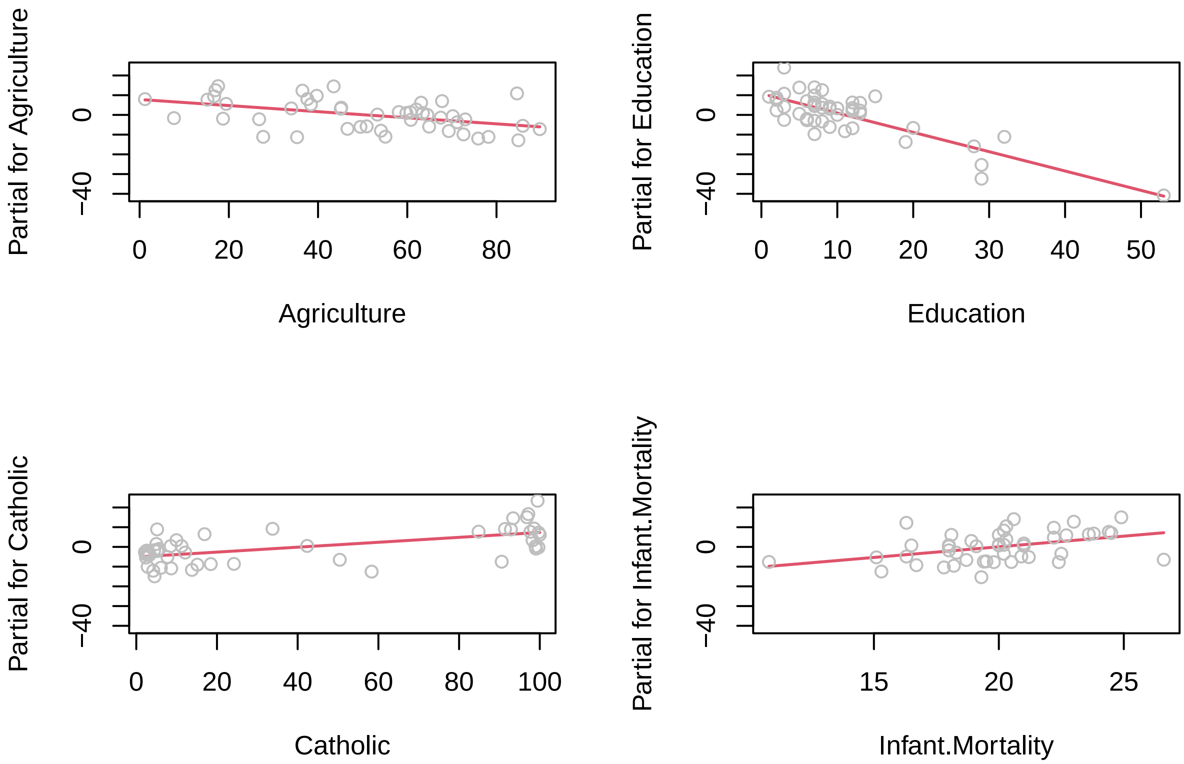
- An exploration of transformations to improve the fit of the model.

```
plmodi <- lm(Fertility ~ Agriculture + Education + Catholic + Infant.Mortality,
             data = swiss)
# summary(plmodi)
plmodi %>%
  tbl_regression() %>%
  bold_labels() %>%
  bold_p(t = 0.05)
```

```
## Table printed with ‘knitr::kable()‘, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include ‘message = FALSE‘ in code chunk header.
```

Characteristic	Beta	95% CI	p-value
Agriculture	-0.15	-0.29, -0.02	**0.029**
Education	-0.98	-1.3, -0.68	**<0.001**
Catholic	0.12	0.07, 0.18	**<0.001**
Infant.Mortality	1.1	0.31, 1.8	**0.007**

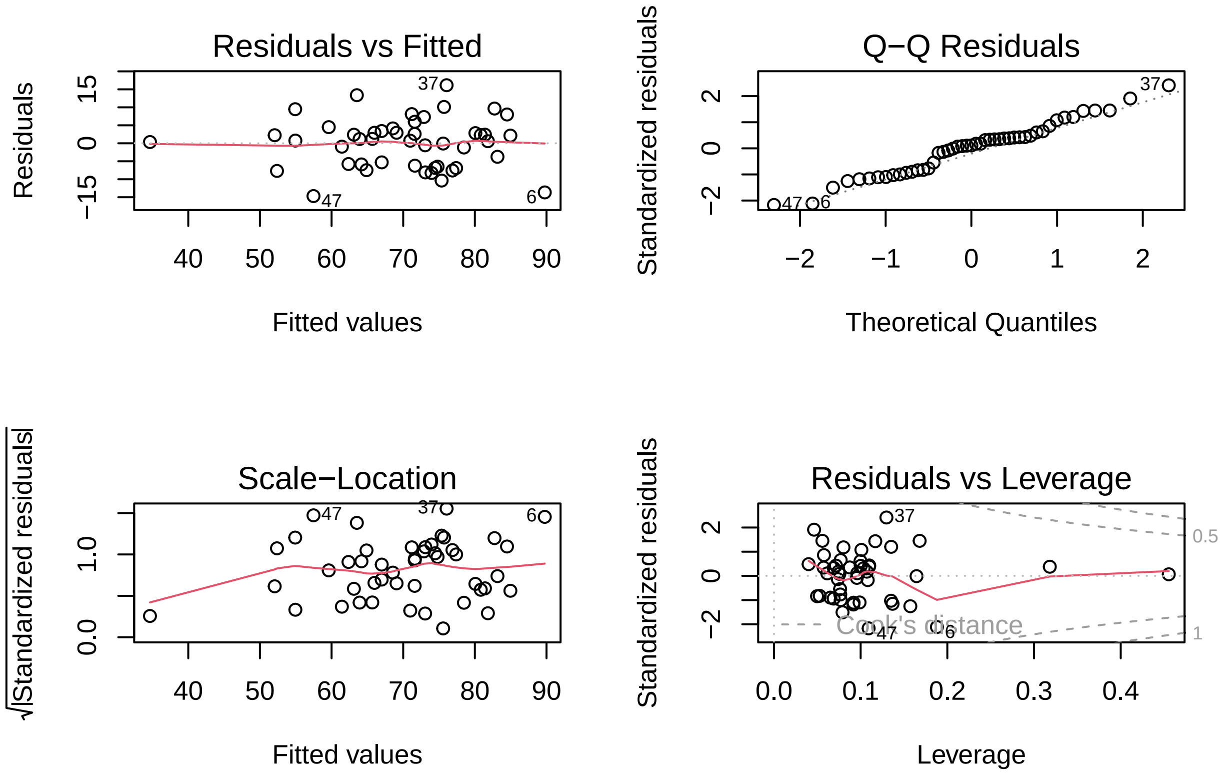
```r
par(mfrow = c(2, 2))
termplot(plmodi, partial.resid = TRUE, terms = NULL)
```



Answer: By observing the partial residual plot, We cannot see there is non-linear relationship between the response and the predictors. Therefore, transformation is unnecessary.

- Diagnostics to check the assumptions of your model.

```r
par(mfrow = c(2, 2))
plot(plmodi)
```



Answer: Based on the diagnostic plots, the residuals seem to satisfy our assumptions for linear regression model. No influentual points are observed in `Residuals vs Leverage` plot.

- Some predictions of future observations for interesting values of the predictors.

```
newdata <- data.frame(Agriculture = 50, Education = 10, Catholic = 10,
                      Infant.Mortality = 19)
predict(plmodi, newdata = newdata, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 66.30486 51.58128 81.02845
```

```
newdata2 <- data.frame(Agriculture = 50, Education = 10, Catholic = 15,
                       Infant.Mortality = 19)
predict(plmodi, newdata = newdata2, interval = "prediction")
```

```
##       fit      lwr     upr
## 1 66.9282 52.23219 81.6242
```

```
newdata3 <- data.frame(Agriculture = 50, Education = 10, Catholic = 20,
                       Infant.Mortality = 19)
predict(plmodi, newdata = newdata3, interval = "prediction")
```

```
##        fit      lwr     upr
## 1 67.55153 52.87736 82.2257
```

```r
newdata4 <- data.frame(Agriculture = 50, Education = 10, Catholic = 30,
                       Infant.Mortality = 19)
predict(plmodi, newdata = newdata4, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 68.79819 54.15037 83.44602
```

```r
newdata5 <- data.frame(Agriculture = 50, Education = 10, Catholic = 40,
                       Infant.Mortality = 19)
predict(plmodi, newdata = newdata5, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 70.04486 55.40019 84.68952
```

Answer: By holding other predictors constant, we modify percentage of catholic to see change of the fertility measure. We observe that under our specific setting of other predictors, the fertility rate seems to be positively associated with percentage of catholic. This might be an indication that countries with more percentage of catholic tends to have higher fertility measure. However, since we only observe the partial effect of the percentage of catholic on the fertility measure, we cannot conclude that the percentage of catholic is directly increase the fertility measure.

- An interpretation of the meaning of the model by writing a scientific abstract. (<150 words)
  - BACKGROUND: brief intro of the study background, what are the existing findings
  - OBJECTIVE: state the overall purpose of your research, e.g., what kind of knowledge gap you are trying to fill in
  - METHODS: study design (how these data were collected), outcome definitions, statistical procedures used
  - RESULTS: summary of major findings to address the question raised in objective
  - CONCLUSIONS:

In 1888, the fertility of Switzerlan was beginning to fall. Previous study suggests that industrialization and economic change might be directly affecting the fertility measure in Europe. In this study, we aims to investigate the relationship between fertility and several predictors including `Agriculture`, `Education`, `Catholic`, and `Infant Mortality`. These predictors are rarely being considered by previous investigators but they are strongly correlated with factor like `industrialization`. Therefore, it worth exploring how these predictors are associated with fertility measure. Data was collected from 47 French-speaking provinces at about 1888. We will will conduct numerical and graphical analysis, model diagnosis, and feature selection before fitting a multiple linear regression model. We found that the percentage of catholic is positively associated with the fertility measure by adjusting other predictors. However, we cannot conclude that the percentage of catholic directly increases the fertility measure in a given country. Further study is needed to explore the underlying mechanism of the relationship between the percentage of catholic and fertility measure.

Q2. Concavity of logistic regression log-likelihood

Q2.1

Write down the log-likelihood function of logistic regression for binomial responses.

Answer:

The logistic regression has the form $p = \frac{e^{X^t \beta}}{1+e^{X^T \beta}}$

The log likelihood function is

$l(\theta) = \sum log[p_i^{y_i}(1-p_i)^{1-y_i}] = \sum_i [y_i \cdot x_i^T \beta - log(1+e^{x_i^T \beta})]$

Q2.2

Derive the gradient vector and Hessian matrix of the log-likelihood function with respect to the regression coefficients $\boldsymbol{\beta}$.

Answer:

The gradient vector is

$$\phi(\theta) = \frac{\partial l(\theta)}{\partial \beta} = \sum_i [y_i \cdot x_i^T - \frac{e^{x_i^T \beta} x_i^T}{1+e^{x_i^T \beta}}] = \sum_i [y_i - \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}](x_i^T)^T$$

Since the first term is a scalar, we can move x_i^T to the right. Since x_i^T indicates the first row vector, we can transform it to represent it as a column vector.

$$D_\beta \phi(\theta) = \sum_i [-(x_i^T)^T D_\beta \frac{e^{x_i^T \beta}}{1+e^{x_i^T \beta}}] = \sum_i [-(x_i^T)^T (D_m \frac{m}{1+m})(D_t e^t)(D_\beta x_i^T \beta)] = \sum_i [-(x_i^T)^T (\frac{1}{(1+m)^2})(e^t)(x_i^T)] = \sum_i [-(x_i^T$$

Q2.3

Show that the log-likelihood function of logistic regression is a concave function in regression coefficients $\boldsymbol{\beta}$. (Hint: show that the negative Hessian is a positive semidefinite matrix.)

Answer:

For all vectors a,

$$\sum_i [a^T (x_i^T)^T (\frac{1}{(1+e^{x_i^T \beta})^2})(e^{x_i^T \beta})(x_i^T) a] = \sum_i [(x_i^T a)^T (\frac{e^{x_i^T \beta}}{(1+e^{x_i^T \beta})^2})(x_i^T a)] = \sum_i [(x_i^T a)^2 (\frac{e^{x_i^T \beta}}{(1+e^{x_i^T \beta})^2})$$

Since exponential function is always positive, the Hessian matrix is positive semidefinite by Energy-based definition. Therefore, the log-likelihood function is concave.

Q3.

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The data may be found in the the dataset `pima`.

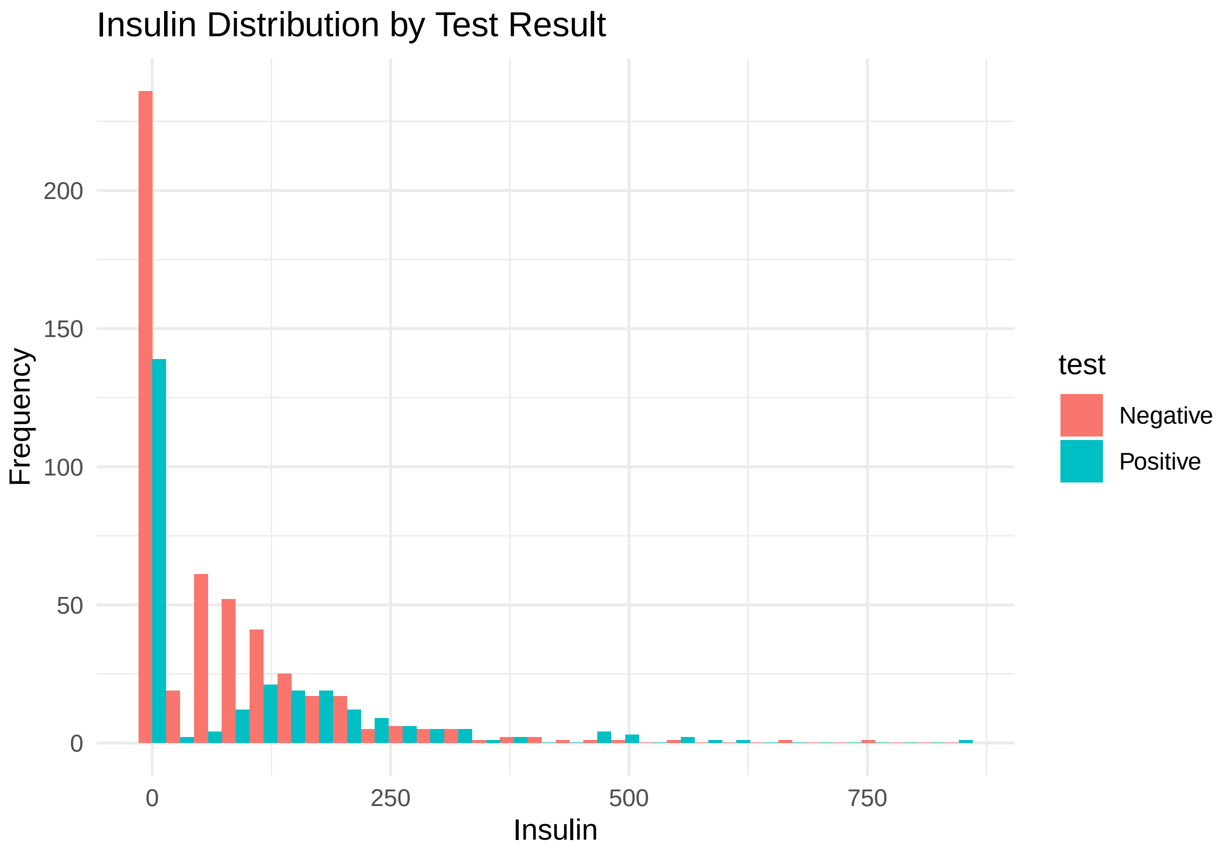
```r
data(pima)
```

Q3.1

Create a factor version of the test results and use this to produce an interleaved histogram to show how the distribution of insulin differs between those testing positive and negative. Do you notice anything unbelievable about the plot?

```r
pima$test <- factor(pima$test, levels = c(0, 1), labels = c("Negative", "Positive"))
```

```r
ggplot(pima, aes(insulin, fill = test)) +
  geom_histogram(position = "dodge", bins = 30) +
  labs(title = "Insulin Distribution by Test Result",
       x = "Insulin",
       y = "Frequency") +
  theme_minimal()
```



Answer: The plot shows that there are many zero values in the insulin variable. This is not possible since insulin is a hormone that is always present in the body. The zero values are likely to be missing values that have been coded as zero. There are also some extremely large insulin values that are likely to be errors. These will need to be investigated further.

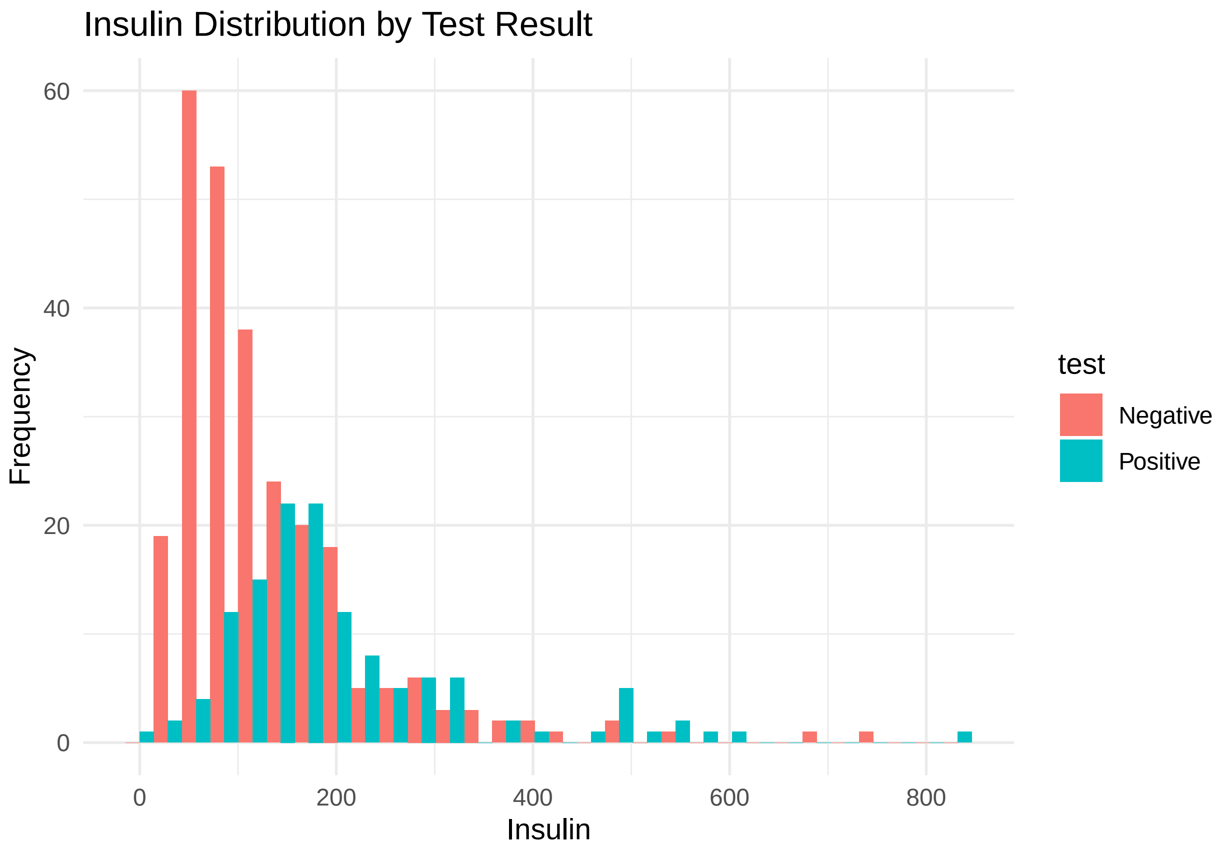
Q3.2

Replace the zero values of `insulin` with the missing value code `NA`. Recreate the interleaved histogram plot and comment on the distribution.

```
pima$insulin[pima$insulin == 0] <- NA
```

```
ggplot(pima, aes(insulin, fill = test)) +
  geom_histogram(position = "dodge", bins = 30) +
  labs(title = "Insulin Distribution by Test Result",
       x = "Insulin",
       y = "Frequency") +
  theme_minimal()
```

```
## Warning: Removed 374 rows containing non-finite values (‘stat_bin()‘).
```



Answer: The plot shows that the distribution of insulin values is different between those testing positive and negative. For those who test negative, they tends to have relatively lower insulin values. However, there are some overlaps of bins between the two groups. It might indicate that we have larger sample size for negative test results.

Q3.3

Replace the incredible zeroes in other variables with the missing value code. Fit a model with the result of the diabetes test as the response and all the other variables as predictors. How many observations were used in the model fitting? Why is this less than the number of observations in the data frame.

```r
pima$glucose[pima$glucose == 0] <- NA
pima$diastolic[pima$diastolic == 0] <- NA
pima$triceps[pima$triceps == 0] <- NA
pima$bmi[pima$bmi == 0] <- NA
pima$age[pima$age == 0] <- NA
```

```r
fit <- glm(test ~ ., data = pima, family = binomial)
```

```r
summary(fit)
```

```
##
## Call:
## glm(formula = test ~ ., family = binomial, data = pima)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.004e+01  1.218e+00  -8.246  < 2e-16 ***
## pregnant     8.216e-02  5.543e-02   1.482  0.13825
## glucose      3.827e-02  5.768e-03   6.635 3.24e-11 ***
## diastolic   -1.420e-03  1.183e-02  -0.120  0.90446
## triceps      1.122e-02  1.708e-02   0.657  0.51128
## insulin     -8.253e-04  1.306e-03  -0.632  0.52757
## bmi          7.054e-02  2.734e-02   2.580  0.00989 **
## diabetes     1.141e+00  4.274e-01   2.669  0.00760 **
## age          3.395e-02  1.838e-02   1.847  0.06474 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.02  on 383  degrees of freedom
##   (376 observations deleted due to missingness)
## AIC: 362.02
##
## Number of Fisher Scoring iterations: 5
```

```r
nrow(pima)-376
```

```
## [1] 392
```

Answer: The model used 392 observations in the fitting. The missing values in the data frame are not used in the fitting. The number of observations used in the fitting is less than the number of observations in the data frame because the missing values are not used in the fitting.

Q3.4

Refit the model but now without the insulin and triceps predictors. How many observations were used in fitting this model? Devise a test to compare this model with that in the previous question.

```r
fit2 <- glm(test ~ . - insulin - triceps, data = pima, family = binomial)
```

```r
summary(fit2)
```

```
##
## Call:
## glm(formula = test ~ . - insulin - triceps, family = binomial,
##     data = pima)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.9604796  1.1818764  -8.428  < 2e-16 ***
## pregnant     0.0840497  0.0550728   1.526 0.126971
## glucose      0.0364863  0.0049973   7.301 2.85e-13 ***
## diastolic   -0.0008002  0.0118034  -0.068 0.945949
## bmi          0.0785728  0.0215674   3.643 0.000269 ***
## diabetes     1.1492368  0.4250340   2.704 0.006854 **
## age          0.0346079  0.0181919   1.902 0.057121 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.88  on 385  degrees of freedom
##   (376 observations deleted due to missingness)
## AIC: 358.88
##
## Number of Fisher Scoring iterations: 5
```

```r
nrow(pima)-376
```

```
## [1] 392
```

```r
anova(fit, fit2, test = "Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: test ~ pregnant + glucose + diastolic + triceps + insulin + bmi +
##     diabetes + age
## Model 2: test ~ (pregnant + glucose + diastolic + triceps + insulin +
##     bmi + diabetes + age) - insulin - triceps
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       383     344.02
## 2       385     344.88 -2 -0.85931   0.6507
```

Answer: The model used 392 observations in the fitting. We conduct analysis of deviance with `anova` to test whether model 1 is superior than model 2. The result suggests that there is not enough evidence to reject the null hypothesis, which implies that the two models do not differ significantly in terms of their fit to the data.

Q3.5

Use AIC to select a model. You will need to take account of the missing values. Which predictors are selected? How many cases are used in your selected model?

```r
summary(is.na(pima))
```

```
##   pregnant        glucose        diastolic        triceps
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:768       FALSE:763       FALSE:733       FALSE:541
##                  TRUE :5         TRUE :35        TRUE :227
##   insulin           bmi          diabetes          age
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:394       FALSE:757       FALSE:768       FALSE:768
##  TRUE :374       TRUE :11
##     test
##  Mode :logical
##  FALSE:768
##
```

```r
pimadropna <- pima %>%
  na.omit()
```

```r
summary(is.na(pimadropna))
```

```
##   pregnant        glucose        diastolic        triceps
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:392       FALSE:392       FALSE:392       FALSE:392
##   insulin           bmi          diabetes          age
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical
##  FALSE:392       FALSE:392       FALSE:392       FALSE:392
##     test
##  Mode :logical
##  FALSE:392
```

```r
biglm <- glm(test ~ ., data = pimadropna, family = binomial)
summary(biglm)
```

```
##
## Call:
## glm(formula = test ~ ., family = binomial, data = pimadropna)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.004e+01  1.218e+00  -8.246  < 2e-16 ***
## pregnant     8.216e-02  5.543e-02   1.482  0.13825
## glucose      3.827e-02  5.768e-03   6.635 3.24e-11 ***
## diastolic   -1.420e-03  1.183e-02  -0.120  0.90446
## triceps      1.122e-02  1.708e-02   0.657  0.51128
## insulin     -8.253e-04  1.306e-03  -0.632  0.52757
## bmi          7.054e-02  2.734e-02   2.580  0.00989 **
## diabetes     1.141e+00  4.274e-01   2.669  0.00760 **
## age          3.395e-02  1.838e-02   1.847  0.06474 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.02  on 383  degrees of freedom
## AIC: 362.02
##
## Number of Fisher Scoring iterations: 5
```

```
biglm <- glm(test ~ ., data = pimadropna, family = binomial)

step_model <- step(biglm, direction = "back", trace = TRUE)
```

```
## Start:  AIC=362.02
## test ~ pregnant + glucose + diastolic + triceps + insulin + bmi +
##     diabetes + age
##
##             Df Deviance    AIC
## - diastolic  1   344.04 360.04
## - insulin    1   344.42 360.42
## - triceps    1   344.45 360.45
## <none>           344.02 362.02
## - pregnant   1   346.24 362.24
## - age        1   347.55 363.55
## - bmi        1   350.89 366.89
## - diabetes   1   351.58 367.58
## - glucose    1   396.95 412.95
##
## Step:  AIC=360.04
## test ~ pregnant + glucose + triceps + insulin + bmi + diabetes +
##     age
##
##            Df Deviance    AIC
## - insulin   1   344.42 358.42
## - triceps   1   344.46 358.46
## <none>          344.04 360.04
## - pregnant  1   346.24 360.24
## - age       1   347.60 361.60
## - bmi       1   351.28 365.28
## - diabetes  1   351.67 365.67
## - glucose   1   397.31 411.31
##
## Step:  AIC=358.42
## test ~ pregnant + glucose + triceps + bmi + diabetes + age
##
##            Df Deviance    AIC
## - triceps   1   344.89 356.89
## <none>          344.42 358.42
## - pregnant  1   346.74 358.74
## - age       1   347.87 359.87
## - bmi       1   351.32 363.32
## - diabetes  1   351.90 363.90
## - glucose   1   411.11 423.11
##
## Step:  AIC=356.89
## test ~ pregnant + glucose + bmi + diabetes + age
##
##            Df Deviance    AIC
## <none>          344.89 356.89
## - pregnant  1   347.23 357.23
## - age       1   348.72 358.72
## - diabetes  1   352.72 362.72
## - bmi       1   360.44 370.44
## - glucose   1   411.85 421.85
```

```r
summary(step_model)
```

```
##
## Call:
## glm(formula = test ~ pregnant + glucose + bmi + diabetes + age,
##     family = binomial, data = pimadropna)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.992080   1.086866  -9.193  < 2e-16 ***
## pregnant     0.083953   0.055031   1.526 0.127117
## glucose      0.036458   0.004978   7.324 2.41e-13 ***
## bmi          0.078139   0.020605   3.792 0.000149 ***
## diabetes     1.150913   0.424242   2.713 0.006670 **
## age          0.034360   0.017810   1.929 0.053692 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 498.10  on 391  degrees of freedom
## Residual deviance: 344.89  on 386  degrees of freedom
## AIC: 356.89
##
## Number of Fisher Scoring iterations: 5
```

Answer: Rows containing missing values are dropped. `5` predictors are selected based AIC in `backward selection`. The selected predictors are `pregnant`, `glucose`, `bmi`, `diabetes`, and `age`. The model used `392` observations in the fitting.

Q3.6

Create a variable that indicates whether the case contains a missing value. Use this variable as a predictor of the test result. Is missingness associated with the test result? Refit the selected model, but now using as much of the data as reasonable. Explain why it is appropriate to do this.

```r
library(faraway)
library(tidyverse)
```

```r
pima <- pima %>%
  mutate(
    glucose2  = ifelse(glucose == 0, NA, glucose),
    diastolic2 = ifelse(diastolic == 0, NA, diastolic),
    triceps2 = ifelse(triceps == 0, NA, triceps),
    insulin2 = ifelse(insulin == 0, NA, insulin),
    bmi2 = ifelse(bmi == 0, NA, bmi),
    diabetes2 = ifelse(diabetes == 0, NA, diabetes),
    age2 = ifelse(age == 0, NA, age))

pima$missingNA = ifelse(apply(is.na(dplyr::select(pima, contains("2"))), 1, sum) > 0, 1, 0)

missing.glm <- glm(test ~ missingNA, family = binomial(), data = pima)

library(gtsummary)
missing.glm %>%
  tbl_regression() %>%
  bold_labels() %>%
  bold_p(t = 0.05)
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**log(OR)**	**95% CI**	**p-value**
missingNA	0.16	-0.14, 0.45	0.3

From above regression, we found missingness is not significantly associated with outcome since p is greater than 0.05. This means that the distribution of outcome when removing data with missing is still a representative of the original distribution. This justifies the use of "complete case" analysis.

```r
library(dplyr)
pimaSelected <- pima |>
  collect() |>
  dplyr::select(test, pregnant, glucose, bmi, diabetes, age)
```

```r
pimaSelected <- pimaSelected |>
  na.omit()
```

```r
refitlm <- glm(test ~ ., data = pimaSelected, family = binomial)
```

```r
refitlm
```

```
##
## Call:  glm(formula = test ~ ., family = binomial, data = pimaSelected)
##
## Coefficients:
## (Intercept)     pregnant      glucose          bmi     diabetes          age
##    -9.32279      0.11506      0.03594      0.08753      0.92058      0.01137
```

```
##
## Degrees of Freedom: 751 Total (i.e. Null);  746 Residual
## Null Deviance:       974.7
## Residual Deviance: 703.2     AIC: 715.2
```

Answer: This is appropriate because missingness is not significantly associated with the test result. The selected model is refitted using the complete cases. This can give us more information and more power in hypothesis testing.

Q3.7

Using the last fitted model of the previous question, what is the odd ratio of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant? Give a confidence interval for this difference.

```
# Calculate the first and third quartiles for BMI
bmi_q1 <- quantile(pima$bmi, 0.25, na.rm = TRUE)
bmi_q3 <- quantile(pima$bmi, 0.75, na.rm = TRUE)

bmi_diff = bmi_q1 - bmi_q3
```

```
summary(refitlm)
```

```
##
## Call:
## glm(formula = test ~ ., family = binomial, data = pimaSelected)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.322789   0.737279 -12.645  < 2e-16 ***
## pregnant     0.115058   0.032341   3.558 0.000374 ***
## glucose      0.035941   0.003555  10.110  < 2e-16 ***
## bmi          0.087529   0.014722   5.945 2.76e-09 ***
## diabetes     0.920583   0.300832   3.060 0.002212 **
## age          0.011366   0.009315   1.220 0.222405
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 974.75  on 751  degrees of freedom
## Residual deviance: 703.24  on 746  degrees of freedom
## AIC: 715.24
##
## Number of Fisher Scoring iterations: 5
```

```
bmi_coef = coef(refitlm)['bmi']
```

```
odds_ratio <- exp(bmi_coef * bmi_diff)

# Calculate standard error of the BMI coefficient
```

```
se_bmi <- summary(refitlm)$coefficients["bmi", "Std. Error"]

# Z-value for 95% confidence; approximately 1.96 for 95% CI
z_value <- 1.96

# Log-odds interval
log_odds_low <- log(odds_ratio) - z_value * se_bmi * bmi_diff
log_odds_high <- log(odds_ratio) + z_value * se_bmi * bmi_diff

# Convert log-odds interval back to odds ratio
ci_low <- exp(log_odds_low)
ci_high <- exp(log_odds_high)

# Output the confidence interval
cat("Odds Ratio:", odds_ratio, "\n")
```

```
## Odds Ratio: 0.4508975
```

```
cat("95% Confidence Interval for Odds Ratio: [", ci_low, ", ", ci_high, "]\n")
```

```
## 95% Confidence Interval for Odds Ratio: [ 0.5862989 ,  0.3467661 ]
```

Answer: The odds ratio of testing positive for diabetes for a woman with a BMI at the first quartile compared with a woman at the third quartile, assuming that all other factors are held constant is `0.4508975`. The 95% confidence interval for the odds ratio is `[0.5862989, 0.3467661]`.

Q3.8

Do women who test positive have higher diastolic blood pressures? Is the diastolic blood pressure significant in the regression model? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

Answer:

For question 1, it did not specify who we are comparing with. I assume we are comparing women who test positive have higher diastolic blood pressures than women who test negative. Under this assumption, we are comparing 2 groups. We can use a paired t test to compare the diastolic blood pressure.

For question 2, it is asking if the covariate `diastolic blood pressure` is a significant predictor in the regression model. We can conduct F test to test the significance of the covariate.

The answers are only apparently contradictory because the paired t test is comparing the mean of 2 groups, while the F test is testing the significance of the covariate in the regression model. General speaking, question 1 is comparing between 2 groups, while question 2 is testing the significance of one covariate in the regression model.