Homework 2 - Question 2

Nickhil Sethi

Set Up

Use the swiss data with Fertility as the response:

```
data("swiss")
# commenting this out because
# the output takes up way too much space
# ?swiss
```

Directed Analysis

Use Agriculture as the only predictor

```
lmod <- lm(Fertility ~ Agriculture, data = swiss)
```

1. What are the estimates for $beta_0$ and $beta_1$? Show that the `lm` function gives the same output as a calculation by hand.

Let's print the coefficients from our model:

```
coef(lmod)
```

```
## (Intercept) Agriculture
##  60.3043752   0.1942017
```

Now we run a manual estimate and notice that it gives us nearly the exact same coefficients:

```
# now the manual estimates
manual_estimate_beta_1 <- cov(swiss$Fertility, swiss$Agriculture) / var(swiss$Agriculture)
manual_estimate_beta_0 <- mean(swiss$Fertility) - manual_estimate_beta_1 * mean(swiss$Agriculture)

# notice they're the same as above
manual_estimate_beta_1
```

```
## [1] 0.1942017
```

```
manual_estimate_beta_0
```

```
## [1] 60.30438
```

2. Interpret the results - what do each of the coefficients represent? Are there any considerations about the range of values that the model is applicable for?

The coefficient β_0 represents the expected fertility of a town with 0% men in the agriculture workforce, and β_1 represents the change in percent fertility associated with a 1 percent increase in the men with agriculture as occupation.

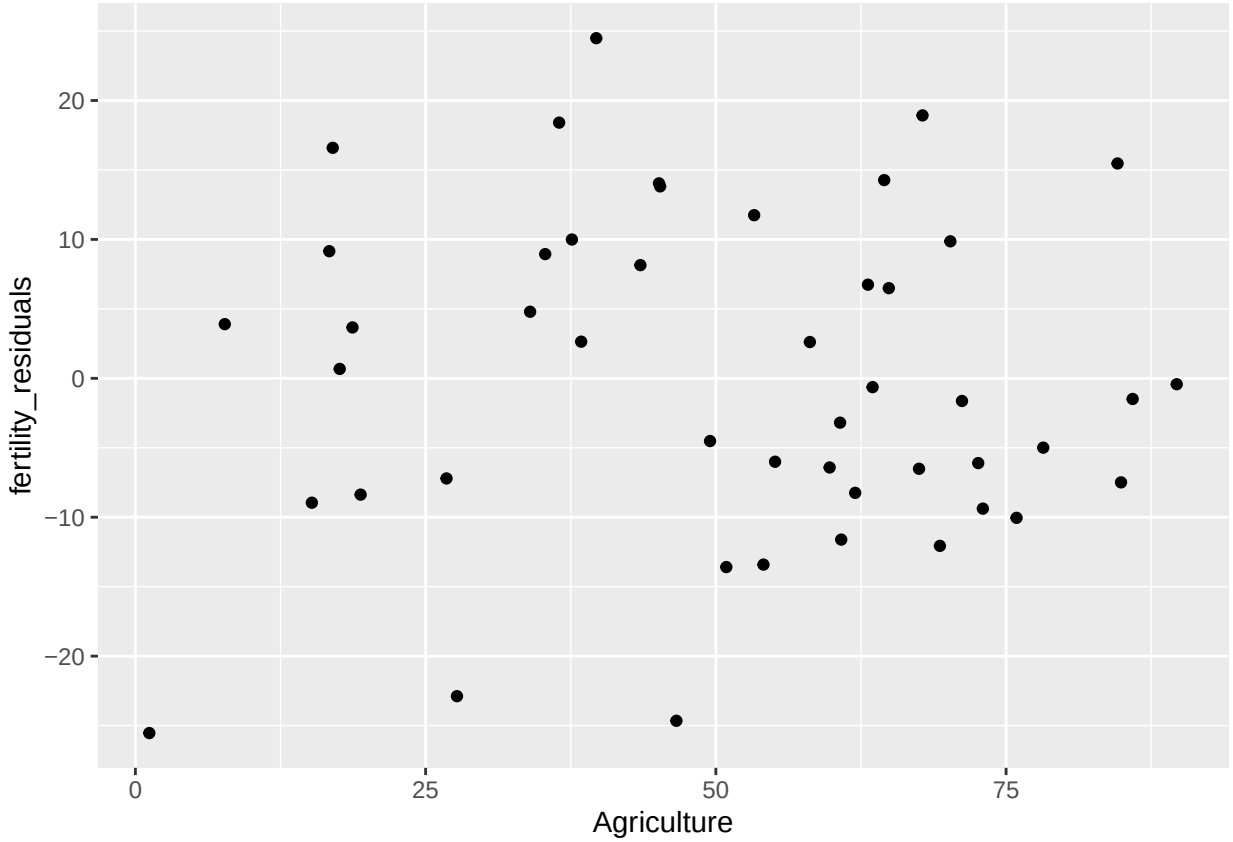
As for the range of values the model applies to – because `swiss$Agriculture` is a percentage, the coefficient β_1 is only applicable to input values between 0 and 1. Likewise, as the response is a percent as well, the

model parameters are invalid for inputs that produce predictions outside $[0, 100]$

3. Create a few residual plots of your choice and comment on your findings

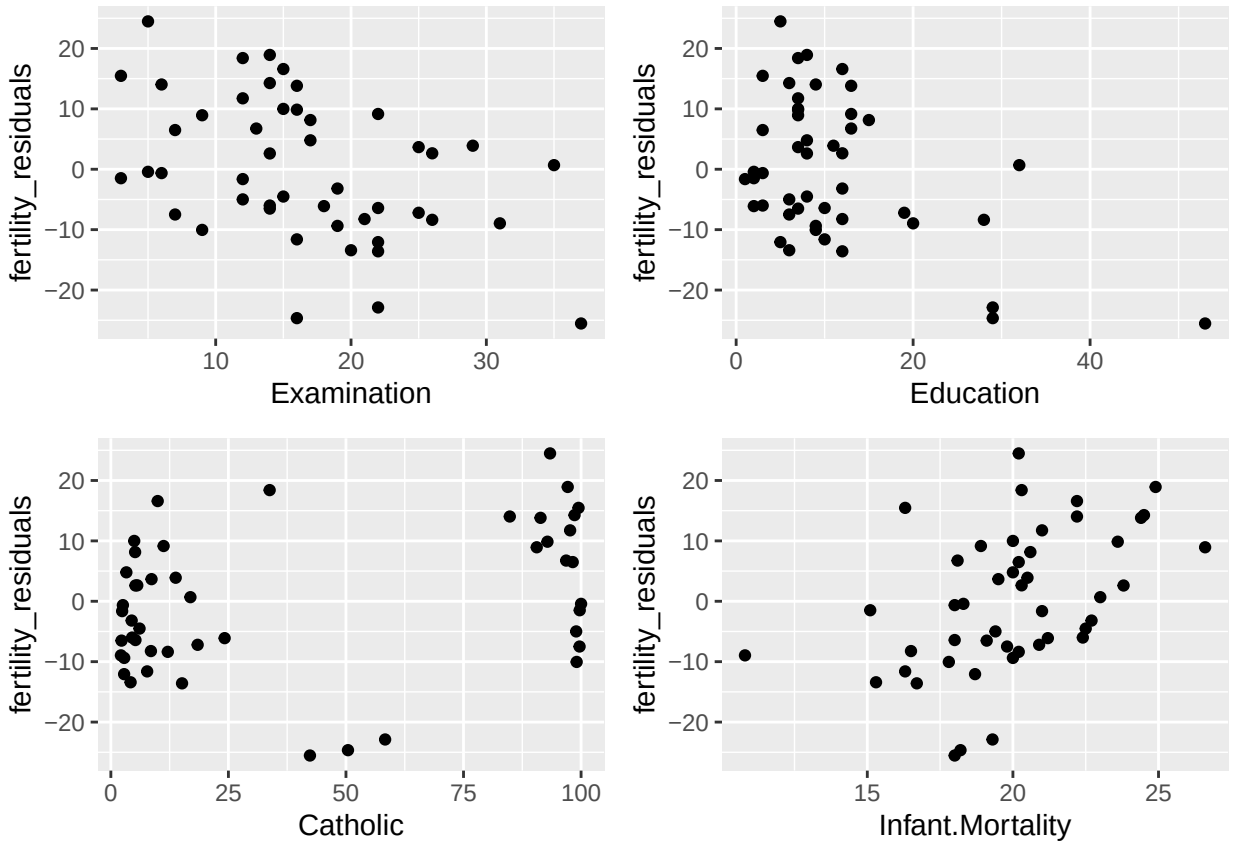
Let's plot the residuals against the other variables

```
fertility_residuals <- residuals(lmod)
p1 <- ggplot(swiss, aes(Agriculture, fertility_residuals)) +
  geom_point()
p2 <- ggplot(swiss, aes(Examination, fertility_residuals)) +
  geom_point()
p3 <- ggplot(swiss, aes(Education, fertility_residuals)) +
  geom_point()
p4 <- ggplot(swiss, aes(Catholic, fertility_residuals)) +
  geom_point()
p5 <- ggplot(swiss, aes(Infant.Mortality, fertility_residuals)) +
  geom_point()
grid.arrange(p1)
```



First, let's take a look at the residuals plotted against the predictor *Agriculture*; these appear to be mean zero and uncorrelated with the predictor, which is consistent with the model assumptions.

```
grid.arrange(p2, p3, p4, p5, ncol=2)
```

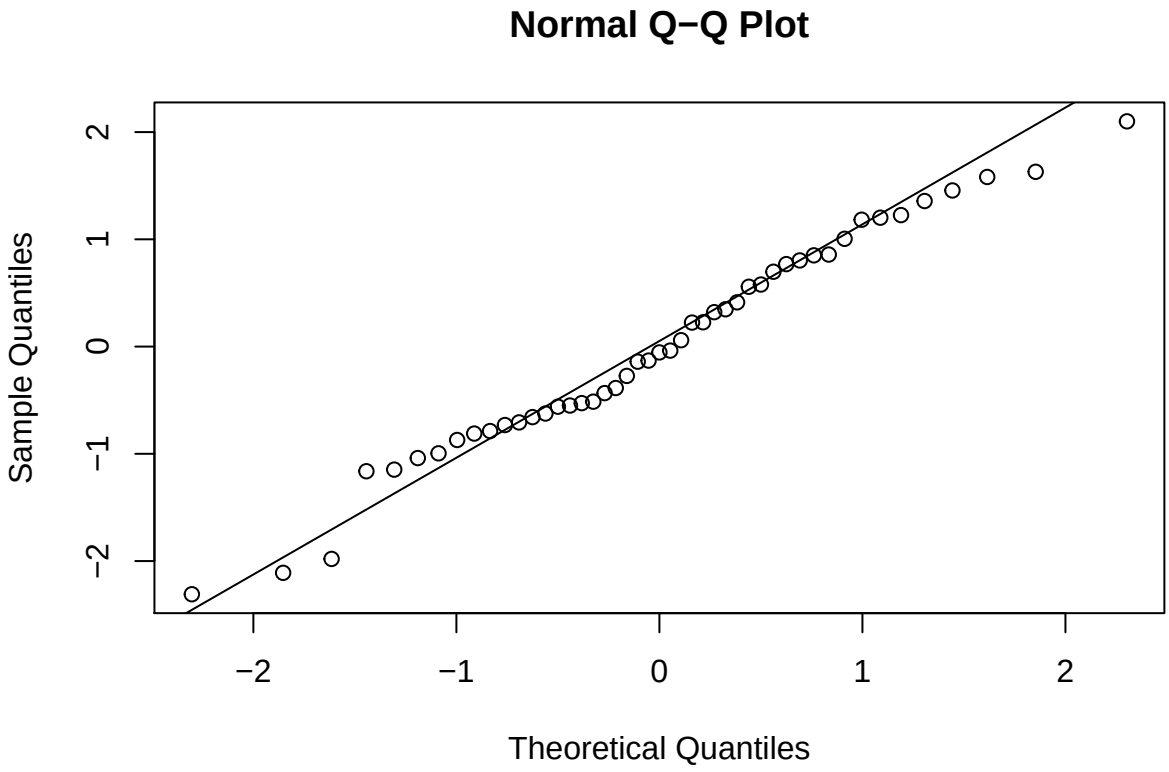


However, the above plots all show a non-trivial relationship between the residuals and the other dimensions of the data set. *Examination* and *Education* show negative correlations and $Infant.Mortality$ shows a positive correlation with the resituals.

The relationship between the *Catholic* variable and the residuals is interesting – catholicism appears to be bi-modal, with towns in general being nearly all catholic or nearly all protestant; for the three towns that are mixed, our model seriously underestimates *Fertility*. Given that this is only for three towns, I wouldn't say this is worth including in a larger model.

Finally, let's show a qqplot of the residuals to get a sense of whether or not the normality assumption is correct. It seems it is:

```
qqnorm(rstandard(lmod))
qqline(rstandard(lmod))
```



4. Let's say a French-speaking province was previously left out (at random) that has an 70% of males involved in agriculture as an occupation. What do you exepect the fertility measure to be? Create an 89% interval. Should a confidence or prediction interval be used?

A prediction interval should be used here, because we're interested in the response variable in this case, not the estimated coefficients.

```
# creating the prediction interval
sigma_squared = sum(fertility_residuals * fertility_residuals) / (length(fertility_residuals) - 2)

# setting our parameters
alpha <- .89
new_agriculture <- 70
mean_agriculture <- mean(swiss$Agriculture)
squared_dev <- (new_agriculture - mean_agriculture)**2
n <- length(fertility_residuals)

x_sum_of_squared_devs <- sum(
  (swiss$Agriculture - mean_agriculture)**2)

sigma_squared_y_hat <- sigma_squared * (
    1 + (1/n) + (squared_dev / x_sum_of_squared_devs ))
sigma_y_hat = sqrt(sigma_squared_y_hat)

new.df <- data.frame(
  Agriculture=c(new_agriculture))

# this gives us the mean of the
# predictive distribution
y_hat <- predict(
```

```
  lmod, newdata = new.df, interval="predict")

pred_interval_upper <- y_hat[1] + sigma_y_hat*qt((1 + alpha)/2, df = n-2)
pred_interval_lower <- y_hat[1] + sigma_y_hat*qt((1 - alpha)/2, df = n-2)

# our lower and upper bounds
# of the .89 prediction interval:
pred_interval_lower
```

```
## [1] 54.27978
```

```
pred_interval_upper
```

```
## [1] 93.51721
```

Analyses of your choice

Pick 2 other models (with different variables included) and compare them to the previous analysis. Which fits the data better? Convince me that you have picked the best model using some of the tools we learned about in class and that are covered in the Introduction chapter.

```
# here's the AIC of our previous model
AIC(lmod)
```

```
## [1] 369.4675
```

```
# now let's include Examination Infant.Mortality and Education
lmod2 <- lm(
  Fertility ~ Agriculture + Examination + Infant.Mortality + Education,
  data = swiss)
AIC(lmod2)
```

```
## [1] 333.1348
```

```
# and finally let's add Catholic as well
lmod3 <- lm(
  Fertility ~ Agriculture + Examination +
    Infant.Mortality + Education + Catholic,
  data = swiss)
AIC(lmod3)
```

```
## [1] 326.0716
```

Let's use *AIC* to examine the use of two new potential models, one which uses all variables and another which uses all but *Catholic*. We see that the *AIC* is actually minimized when we include all the variables; as *AIC* incorporates both the number of variables and the goodness of fit, we can conclude that *lmod*3 is our "best" model.