Name - NetID

Homework 5

Homework Instructions

For questions that require code, please create a code chunk directly below the question and type your code there. Your knitted pdf will show both your code and your output. You are encouraged to knit your file as you work to check that your coding and formatting is done so appropriately.

For written responses or multiple choice questions, please bold your (selected) answer.

Grading Details

All questions will be graded full credit (1 point), half credit (0.5 point) or no credit (0 points).

Full credit responses should have the correct response and appropriate code (if applicable). Half credit responses will have a reasonable attempt (typically no more than one small error or oversight), and no credit responses will be either non-attempts or attempts with significant errors.

Exercise 1

Consider testing for significance of regression in a multiple linear regression model with 9 predictors (and an intercept) and 30 observations. If the value of the F test statistic is 2.4, what is the p-value of this test?

```r
#solution
n = 30
p = 9 + 1
pf(2.4, df1 = p - 1, df2 = n - p, lower.tail = FALSE)
```

```
## [1] 0.04943057
```

Exercise 2

For the next several exercises, use the `swiss` dataset, which is built into `R`. You should use `?swiss` to learn about the background of this dataset.

Fit a multiple linear regression model with `Fertility` as the response and the remaining variables as predictors.

Use your fitted model to make a prediction for the fertility rate of a Swiss province in 1888 with:

- 54% of males involved in agriculture as occupation
- 23% of draftees receiving highest mark on army examination
- 13% of draftees obtaining education beyond primary school
- 60% of the population identifying as Catholic
- 24% of live births that live less than a year

```
#solution
providence = data.frame(
  Agriculture = 54,
  Examination = 23,
  Education = 13,
  Catholic = 60,
  Infant.Mortality = 24
)
swiss_mod = lm(Fertility ~ ., data = swiss)
predict(swiss_mod, newdata = providence)
```

```
##        1
## 72.46069
```

Exercise 3

Create a 99% confidence interval for the coefficient for `Catholic`. Report both the lower and upper bound.

```
#solution
swiss_mod = lm(Fertility ~ ., data = swiss)
confint(swiss_mod, level = 0.99)["Catholic",]
```

```
##       0.5 %      99.5 %
## 0.008877479 0.199353183
```

Exercise 4

Using the summary output, extract the p-value of the test $H_0 : \beta_{\text{Examination}} = 0$ vs $H_1 : \beta_{\text{Examination}} \neq 0$

```
#solution
swiss_mod = lm(Fertility ~ ., data = swiss)
summary(swiss_mod)$coefficients["Examination", "Pr(>|t|)"]
```

```
## [1] 0.3154617
```

Exercise 5

Interpret the p-value reported in the previous exercise. Which is the most appropriate interpretation?

- If Examination *has no* linear relationship with Fertility, we would expect to see at least this much correlation 31.5% of the time.
- **If Examination *has no* linear relationship with Fertility after controlling for these other 4 predictors, we would expect to see at least this much correlation 31.5% of the time.**
- If Examination *does have* a linear relationship with Fertility, we would expect to see at least this much correlation 31.5% of the time.
- If Examination *does have* a linear relationship with Fertility after controlling for these other 4 predictors, we would expect to see at least this much correlation 31.5% of the time.

Exercise 6

Create a 95% confidence interval for the average `Fertility` for a Swiss province in 1888 with:

- 40% of males involved in agriculture as occupation
- 28% of draftees receiving highest mark on army examination
- 10% of draftees obtaining education beyond primary school
- 42% of the population identifying as Catholic
- 27% of live births that live less than a year

Report the **lower** bound of this interval only.

```r
#solution
providence = data.frame(
  Agriculture = 40,
  Examination = 28,
  Education = 10,
  Catholic = 42,
  Infant.Mortality = 27
)
swiss_mod = lm(Fertility ~ ., data = swiss)
predict(swiss_mod, newdata = providence, interval = "confidence")[, "lwr"]
```

```
## [1] 69.4446
```

Exercise 7

Create a 95% prediction interval for the `Fertility` of a Swiss province in 1888 with:

- 40% of males involved in agriculture as occupation
- 28% of draftees receiving highest mark on army examination
- 10% of draftees obtaining education beyond primary school
- 42% of the population identifying as Catholic
- 27% of live births that live less than a year

Yes, these are the same values as for the previous exercise.

Report the **upper** bound of this interval only.

```r
#solution
providence = data.frame(
  Agriculture = 40,
  Examination = 28,
  Education = 10,
  Catholic = 42,
  Infant.Mortality = 27
)
swiss_mod = lm(Fertility ~ ., data = swiss)
predict(swiss_mod, newdata = providence, interval = "prediction")[, "upr"]
```

```
## [1] 94.13635
```

Exercise 8

Run a model summary and observe the F-test result from this model. Determine if this model is explaining a "statistically significant" amount of variance in comparison to the null model. Use $\alpha = 0.01$. What decision do you make?

```
#solution
summary(swiss_mod)
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2743  -5.2617   0.5032   4.1198  15.3213
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *
## Examination      -0.25801    0.25388  -1.016  0.31546
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 **
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

- Fail to reject H_0
- **Reject** H_0
- Reject H_1
- Not enough information

Exercise 9

Consider a model that only uses the predictors `Education`, `Catholic`, and `Infant.Mortality`. Use an F test to compare this with the model that uses all predictors. Run an anova to compare these models and extract the p-value of the resulting F-test.

```
#solution
null_mod = lm(Fertility ~ Education + Catholic + Infant.Mortality, data = swiss)
full_mod = lm(Fertility ~ ., data = swiss)
anova(null_mod, full_mod)[2, "Pr(>F)"]
```

```
## [1] 0.05628314
```

Exercise 10

Consider two nested multiple linear regression models fit to the same data (not necessarily the `swiss` data–in general for any data! One has an R^2 of 0.9 while the other has an R^2 of 0.8. Which model is using fewer predictors?

- The model with an R^2 of 0.9
- **The model with an R^2 of 0.8**
- Not enough information

That is because as predictors are added, R^2 can only go up. So the model with a higher R^2 must be the model with more predictors.

Exercise 11

The following multiple linear regression is fit to data

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

If $\hat{\beta}_1 = 5$ and $\hat{\beta}_2 = 0.25$ then:

- The p-value for testing $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ will be *larger than* the p-value for testing $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$.
- The p-value for testing $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ will be *smaller than* the p-value for testing $H_0 : \beta_2 = 0$ vs $H_1 : \beta_2 \neq 0$.
- **Not enough information**

That is because the coefficient tells us nothing about significance of the test result. Since each predictor is a variable with different units and different amount of variability, we can only compare the standardized t-values of each coefficient, rather than the value of the coefficients themselves.

Exercise 12

Suppose you have a SLR model for predicting IQ from height. The estimated coefficient for height is positive. Now, we add a predictor for age to create a MLR model. After fitting this new model, which result **could** happen to the estimated coefficient?

- Remain exactly the same as the SLR model.
- Be different, but will still positive.
- Become Zero.
- Become Negative.
- **Any of the above.**

Each coefficient is the best estimate for the change in Y given a one unit increase in X_p with all other X holding constant. So if we add a new predictor, then the relationship will likely change and may even change sign… in other words, all options are possible!