Week 4 Quiz Material

When copy and pasting from a code block, or from your local `R` session, be sure to include all available digits for any numeric answer. Also, do not modify the default digits option in the code blocks or your local `R` session.

Practice

Exercise 1

```
#
```

Consider a random variable X that has a F distribution with 3 and 5 degrees of freedom. Calculate $P[X > 2.7]$.

```
1 - pf(2.7, df1 = 3, df2 = 5)
```

```
## [1] 0.1561342
```

```
pf(2.7, df1 = 3, df2 = 5, lower.tail = FALSE)
```

```
## [1] 0.1561342
```

- Hint: By default `pf()` considers area under the curve to the left of the given value.

Exercise 2

```
#
```

For this following Exercises, use the built-in `longley` dataset in `R`. Fit a multiple linear regression model with `Employed` as the response. Use three predictors: `GNP`, `Population`, and `Armed.Forces`. Specifically

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

where

- x_1 is `GNP`
- x_2 is `Population`
- x_3 is `Armed.Forces`

Create a 90% confidence interval for β_1. Report the lower bound of this interval.

```
longley_model = lm(Employed ~ GNP + Population + Armed.Forces, data = longley)
confint(longley_model, level = 0.90, parm = "GNP")[1]
```

```
## [1] 0.05579357
```

- Hint: Include only the specified predictors.

- Hint: Remember to specify the confidence level.

Exercise 3

```
#
```

What is the standard error of $\hat{\beta}_2$?

```
longley_model = lm(Employed ~ GNP + Population + Armed.Forces, data = longley)
summary(longley_model)$coefficients["Population", "Std. Error"]
```

```
## [1] 0.1859156
```

- Hint: Utilize the `summary()` function.

Exercise 4

```
#
```

What is the p-value for testing $H_0 : \beta_3 = 0$ vs $H_1 : \beta_3 \neq 0$?

```
longley_model = lm(Employed ~ GNP + Population + Armed.Forces, data = longley)
summary(longley_model)$coefficients["Armed.Forces", "Pr(>|t|)"]
```

```
## [1] 0.0970466
```

- Hint: Utilize the `summary()` function.

Exercise 5

```
#
```

What is the value of the F test statistic for testing for significance of regression?

```
longley_model = lm(Employed ~ GNP + Population + Armed.Forces, data = longley)
summary(longley_model)$fstatistic["value"]
```

```
##    value
## 238.5757
```

- Hint: Utilize the `summary()` function.

Graded

Exercise 1

```
#
```

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

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

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

Exercise 2

```
#
```

What is the p-value for testing $H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$in a multiple linear regression model with 5 predictors and 20 observations if the value of the t test statistic is -1.3?

```
n = 20
p = 6
2 * pt(abs(-1.3), df = n - p, lower.tail = FALSE)
```

```
## [1] 0.2145976
```

Exercise 3

```
set.seed(42)
x_values = data.frame(
  x1 = runif(15),
  x2 = runif(15),
  x3 = runif(15)
)
```

Consider the true model

$$Y = 3 + 2x_1 + 0.5x_2 + 5x_3 + \epsilon$$

where

$$\epsilon \sim N(0, \sigma^2 = 9)$$

What is $\text{SD}[\hat{\beta}_2]$ given the values of predictors above?

```
X = as.matrix(cbind(rep(1, 15), x_values))
C = solve(t(X) %*% X)
sqrt(9 * C[2 + 1, 2 + 1])
```

```
## [1] 2.47399
```

$$\text{Var}[\hat{\beta}_j] = \sigma^2 C_{jj} \qquad C = \left(X^\top X\right)^{-1}$$

Exercise 4

```
#
```

For exercises 4 - 11, use the `swiss` dataset, which is built into `R` .

Fit a multiple linear regression model with `Fertility` as the response and the remaining variables as predictors. You should use `?swiss` to learn about the background of this dataset.

Use your fitted model to make a prediction for 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

```
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 5

```
#
```

Create a 99% confidence interval for the coefficient for `Catholic` . Report the upper bound of this interval.

```
swiss_mod = lm(Fertility ~ ., data = swiss)
confint(swiss_mod, level = 0.99)["Catholic", "99.5 %"]
```

```
## [1] 0.1993532
```

Exercise 6

```
#
```

Calculate the p-value of the test $H_0 : \beta_{\text{Examination}} = 0$ vs $H_1 : \beta_{\text{Examination}} \neq 0$

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

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

Exercise 7

```
#
```

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.

```
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 8

```
#
```

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

Report the lower bound of this interval.

```
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")[, "lwr"]
```

```
## [1] 60.96392
```

Exercise 9

```
#
```

Report the value of the F statistic for the significance of regression test.

```
swiss_mod = lm(Fertility ~ ., data = swiss)
summary(swiss_mod)$fstatistic["value"]
```

```
##    value
## 19.76106
```

Exercise 10

```
#
```

Carry out the significance of regression test using $\alpha = 0.01$. What decision do you make?

```
null_mod = lm(Fertility ~ 1, data = swiss)
swiss_mod = lm(Fertility ~ ., data = swiss)
anova(null_mod, swiss_mod)[2, "Pr(>F)"] < 0.01
```

```
## [1] TRUE
```

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

Exercise 11

```
#
```

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. Report the p-value of this test.

```
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 12

Consider two nested multiple linear regression models fit to the same data. One has an R^2 of 0.9 while the other has an R^2 of 0.8. Which model uses fewer predictors?

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

Exercise 13

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

Exercise 14

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, the estimated coefficient for height **must be**:

- Exactly the same as the SLR model.
- Different, but still positive.
- Zero.
- Negative.
- **None of the above.**

Exercise 15

The following multiple linear regression is fit to data

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

If the F test for the significance of regression has a p-value less than 0.01, then we know that

- The p-values for both $H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$and $H_0 : \beta_2 = 0 \text{ vs } H_1 : \beta_2 \neq 0$will be less than 0.01.
- **The p-values for both $H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$and $H_0 : \beta_2 = 0 \text{ vs } H_1 : \beta_2 \neq 0$could be greater than 0.01.**
- $H_0 : \beta_1 = 0 \text{ vs } H_1 : \beta_1 \neq 0$will have a p-value less than 0.01 if $H_0 : \beta_2 = 0 \text{ vs } H_1 : \beta_2 \neq 0$has a p-value greater than 0.01.