Week 9 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. It would be best to copy and paste values that were returned using printing methods that do not round results. (Notably the direct output from calling `summary()` .) Also, do not modify the default digits option in the code blocks or your local `R` session.

Practice

Exercise 1

```
# preamble

gen_data = function() {
  n = 50
  x1 = runif(n)
  x2 = runif(n)
  x3 = runif(n)
  x4 = runif(n)
  x5 = x4 + rnorm(n, sd = 0.05)
  x6 = runif(n)
  y = x1 + x3 + x5 + rnorm(n)
  data.frame(y, x1, x2, x3, x4, x5, x6)
}

set.seed(42)
quiz_data = gen_data()
```

```
# starter
quiz_data
```

The above code block has access to a data frame stored in the variable `quiz_data` . We will use `y` as the response, and the remaining variables as predictors. Calculate the partial correlation coefficient between `y` and `x1` controlling for the effect of the remaining variables.

```
# solution
y_mod = lm(y ~ . - x1, data = quiz_data)
x1_mod = lm(x1 ~ x2 + x3 + x4 + x5 + x6, data = quiz_data)
cor(resid(y_mod), resid(x1_mod))
```

```
## [1] 0.2024066
```

- Hint: You will need to obtain the residuals from two models.
- Hint: You will need to use both `y` and `x1` as response variables.

Exercise 2

```
# preamble

gen_data = function() {
  n = 50
  x1 = runif(n)
  x2 = runif(n)
  x3 = runif(n)
  x4 = runif(n)
  x5 = x4 + rnorm(n, sd = 0.05)
  x6 = runif(n)
  y = x1 + x3 + x5 + rnorm(n)
  data.frame(y, x1, x2, x3, x4, x5, x6)
}

set.seed(42)
quiz_data = gen_data()
```

```
# starter
quiz_data
```

The above code block has access to a data frame stored in the variable `quiz_data` . We will use `y` as the response. Fit an additive model using the remaining variables as predictors. Calculate the variance inflation factor of the regression coefficient for `x5` .

```
# solution
x5_mod = lm(x5 ~ x1 + x2 + x3 + x4 + x6, data = quiz_data)
1 / (1 - summary(x5_mod)$r.squared)
```

```
## [1] 39.87626
```

- Hint: Since you might not have access to a `vif()` function since the required packages might not be available, you'll need to use the definition.
- Hint: You'll need to fit a model with `x5` as the response.

Exercise 3

```
# preamble

gen_data = function() {
  n = 50
  x1 = runif(n)
  x2 = runif(n)
  x3 = runif(n)
  x4 = runif(n)
  x5 = x4 + rnorm(n, sd = 0.05)
  x6 = runif(n)
  y = x1 + x3 + x5 + rnorm(n)
  data.frame(y, x1, x2, x3, x4, x5, x6)
}

set.seed(42)
quiz_data = gen_data()
```

```
# starter
quiz_data
```

The above code block has access to a data frame stored in the variable `quiz_data` . We will use `y` as the response. Fit two additive linear models:

- One with all possible predictors.
- One with `x1` , `x2` , and `x3` as predictors.

Use AIC to compare these two models. Report the RSS of the preferred model.

```
# solution
full_mod = lm(y ~ ., data = quiz_data)
smaller_mod = lm(y ~ x1 + x2 + x3, data = quiz_data)

get_rss = function(model) {
  sum(resid(model) ^ 2)
}

ifelse(AIC(full_mod) < AIC(smaller_mod), get_rss(full_mod), get_rss(smaller_mod))
```

```
## [1] 39.66296
```

- Hint: Recall, `R` has built-in functions to compute AIC.
- Hint: Remember, lower is better with AIC.

Exercise 4

```
# preamble

gen_data = function() {
  n = 50
  x1 = runif(n)
  x2 = runif(n)
  x3 = runif(n)
  x4 = runif(n)
  x5 = x4 + rnorm(n, sd = 0.05)
  x6 = runif(n)
  y = x1 + x3 + x5 + rnorm(n)
  data.frame(y, x1, x2, x3, x4, x5, x6)
}

set.seed(42)
quiz_data = gen_data()
```

```
# starter
quiz_data
```

The above code block has access to a data frame stored in the variable `quiz_data`. We will use `y` as the response. Fit two additive linear models:

- One with `x1`, `x2`, and `x4` as predictors.
- One with `x3`, `x4`, `x5`, and `x6` as predictors.

Report the Adjusted R^2 of the model with the better Adjusted R^2.

```
# solution
mod_1 = lm(y ~ x1 + x2 + x4, data = quiz_data)
mod_2 = lm(y ~ x3 + x4 + x5 + x6, data = quiz_data)

max(summary(mod_1)$adj, summary(mod_2)$adj)
```

```
## [1] 0.1390175
```

- Hint: Be sure to report Adjusted R^2, not simply R^2.
- Hint: Remember, higher is better with Adjusted R^2.

Exercise 5

```
# preamble

gen_data = function() {
  n = 50
  x1 = runif(n)
  x2 = runif(n)
  x3 = runif(n)
  x4 = runif(n)
  x5 = x4 + rnorm(n, sd = 0.05)
  x6 = runif(n)
  y = x1 + x3 + x5 + rnorm(n)
  data.frame(y, x1, x2, x3, x4, x5, x6)
}

set.seed(42)
quiz_data = gen_data()
```

```
# starter
quiz_data
```

The above code block has access to a data frame stored in the variable `quiz_data`. We will use `y` as the response. Start with an additive model using the remaining variables as predictors, then perform variable selection using backwards AIC.

Report the LOOCV-RMSE of the chosen mode.

```
# solution
full_model = lm(y ~ ., data = quiz_data)
selected = step(full_model, trace = FALSE)

calc_loocv_rmse = function(model) {
sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2))
}

calc_loocv_rmse(selected)
```

```
## [1] 0.9785782
```

- Hint: Use the `step()` function.
- Hint: Remember, LOOCV-RMSE can be calculated based on a single fit of the regression.

Graded

Exercise 1

For exercises 1 - 9, use the the built-in `R` dataset `mtcars`. Use `mpg` as the response variable. Do not modify any of the data. (An argument could be made for `cyl`, `gear`, and `carb` to be coerced to factors, but for simplicity, we will keep them numeric.)

```
# starter
mtcars
```

Fit an additive linear model with all available variables as predictors. What is the largest variance inflation factor? (Consider answering this question in a local `R` session and use an existing `vif()` function.)

```
# solution
fit = lm(mpg ~ ., data = mtcars)
max(car::vif(fit))
```

```
## [1] 21.62024
```

Exercise 2

```
# starter
mtcars
```

What is the Adjusted R^2 of the model fit in Exercise 1?

```
# solution
fit = lm(mpg ~ ., data = mtcars)
summary(fit)$adj.r.squared
```

```
## [1] 0.8066423
```

Exercise 3

```
# starter
mtcars
```

What is the LOOCV-RMSE of the model fit in Exercise 1?

```
# solution
fit = lm(mpg ~ ., data = mtcars)

calc_loocv_rmse = function(model) {
sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2))
}

calc_loocv_rmse(fit)
```

```
## [1] 3.490209
```

Exercise 4

```
# starter
mtcars
```

Start with the model fit in Exercise 1 then perform variable selection using backwards AIC. Which of the following variables are selected? (Mark all that are selected.)

```
# solution
fit = lm(mpg ~ ., data = mtcars)
selected = step(fit, trace = FALSE)
names(coef(selected))[-1]
```

```
## [1] "wt"   "qsec" "am"
```

- `cyl`
- **`wt`**
- `drat`
- `vs`
- **`qsec`**
- `carb`
- **`am`**

Exercise 5

```
# starter
mtcars
```

What is the LOOCV-RMSE of the model found via selection in Exercise 4?

```
# solution
fit = lm(mpg ~ ., data = mtcars)
selected = step(fit, trace = FALSE)
calc_loocv_rmse(selected)
```

```
## [1] 2.688538
```

Exercise 6

```
# starter
mtcars
```

What is the largest variance inflation factor of the model found via selection in Exercise 4?

```
# solution
fit = lm(mpg ~ ., data = mtcars)
selected = step(fit, trace = FALSE)
max(car::vif(selected))
```

```
## [1] 2.541437
```

Exercise 7

Based on the previous exercises, which of the following is true? (We will refer to the model in Exercise 1 as the "full model" and the model found in Exercise 4 as the "selected model.")

- The selected model is better for predicting, but has collinearity issues.
- The full model is better for predicting, but has collinearity issues.
- **The selected model is better for predicting and does not have collinearity issues.**
- The full model is better for predicting and does not have collinearity issues.

Exercise 8

```
# starter
mtcars
```

Perform variable selection using BIC and a forward search. Begin the search with no predictors. The largest allowable model should be an additive model using all possible predictors.

Which of the following variables are selected? (Mark all that are selected.)

```
# solution
selected = step(lm(mpg ~ 1, data = mtcars),
                mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb,
                k = log(nrow(mtcars)), trace = FALSE)
names(coef(selected))[-1]
```

```
## [1] "wt"  "cyl"
```

- **`wt`**
- `drat`
- **`cyl`**
- `vs`
- `qsec`
- `carb`
- `am`

Exercise 9

```
# starter
mtcars
```

What is the LOOCV-RMSE of the model found via selection in Exercise 8?

```
# solution
selected_bic = step(lm(mpg ~ 1, data = mtcars),
                    mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb,
                    k = log(nrow(mtcars)), trace = FALSE)
calc_loocv_rmse(selected_bic)
```

```
## [1] 2.715962
```

Exercise 10

```
# starter
LifeCycleSavings
```

For exercises 10 - 15, use the the built-in `R` dataset `LifeCycleSavings` . Use `sr` as the response variable.

Calculate the partial correlation coefficient between `sr` and `ddpi` controlling for the effect of the remaining variables.

```
# solution
mod_1 = lm(sr ~ . - ddpi, data = LifeCycleSavings)
mod_2 = lm(ddpi ~ pop15 + pop75 + dpi, data = LifeCycleSavings)
cor(resid(mod_1), resid(mod_2))
```

```
## [1] 0.2972201
```

Exercise 11

```
# starter
LifeCycleSavings
```

Fit a model with all available predictors as well as their two-way interactions. What is the Adjusted R^2 of this model?

```
# solution
fit = lm(sr ~ . ^ 2, data = LifeCycleSavings)
summary(fit)$adj.r.squared
```

```
## [1] 0.261233
```

Exercise 12

```
# starter
LifeCycleSavings
```

Start with the model fit in Exercise 11 then perform variable selection using backwards BIC. Which of the following variables are selected? (Mark all that are selected.)

```
# solution
fit = lm(sr ~ . ^ 2, data = LifeCycleSavings)
selected = step(fit, k = log(nrow(LifeCycleSavings)), trace = FALSE)
names(coef(selected))[-1]
```

```
## [1] "pop15"    "dpi"      "ddpi"     "dpi:ddpi"
```

- `pop15:pop75`
- `pop15:dpi`
- `pop15:ddpi`
- `pop75:dpi`
- `pop75:ddpi`
- **`dpi:ddpi`**

Exercise 13

```
# starter
LifeCycleSavings
```

Start with the model fit in Exercise 11 then perform variable selection using backwards AIC. Which of the following variables are selected? (Mark all that are selected.)

```
# solution
fit = lm(sr ~ . ^ 2, data = LifeCycleSavings)
selected = step(fit, trace = FALSE)
names(coef(selected))[-1]
```

```
## [1] "pop15"    "dpi"      "ddpi"     "dpi:ddpi"
```

- `pop15:pop75`
- `pop15:dpi`
- `pop15:ddpi`
- `pop75:dpi`
- `pop75:ddpi`
- **`dpi:ddpi`**

Exercise 14

```
# starter
LifeCycleSavings
```

Consider the model in Exercise 11, the model found in Exercise 13, and an additive model with all possible predictors. Based of LOOCV-RMSE, which of these models is best? Report the LOOCV-RMSE of the model you choose.

```
# solution
additive = lm(sr ~ ., data = LifeCycleSavings)
fit = lm(sr ~ . ^ 2, data = LifeCycleSavings)
selected = step(fit, trace = FALSE)

min(calc_loocv_rmse(additive),
    calc_loocv_rmse(fit),
    calc_loocv_rmse(selected))
```

```
## [1] 3.833628
```

Exercise 15

```
# starter
LifeCycleSavings
```

Consider the model in Exercise 11, the model found in Exercise 13, and an additive model with all possible predictors. Based of Adjusted R^2, which of these models is best? Report the Adjusted R^2 of the model you choose.

```
# solution
additive = lm(sr ~ ., data = LifeCycleSavings)
fit = lm(sr ~ . ^ 2, data = LifeCycleSavings)
selected = step(fit, trace = FALSE)

max(summary(additive)$adj.r.squared,
    summary(fit)$adj.r.squared,
    summary(selected)$adj.r.squared)
```

```
## [1] 0.3188961
```