Week 1 Quiz Material

When copy and pasting from a code block, or from your local R session, be sure to include all digits for any numeric answer. Also, do no modify the default digits option in the code blocks or your local R session.

Practice

Exercise 1

*# starter*

x = 1:100

Calculate

∑i=1nln(xi).∑i=1nln⁡(xi).

That is, sum the log of each element of x.

Exercise 2

*# starter*

set.seed(42)

a\_vector = rpois(250, lambda = 6)

How many of the elements of a\_vector are greater than or equal to 5? (Notice were using two functions set.seed(), and rpois() to create this vector. We will discuss these at length when we begin to discuss probability.) Be sure to run the two lines in order, otherwise your vector will not contain the expected elements.

Exercise 3

*# starter*

x = 1:100

Create a new vector y, which adds 5 to the elements stored in odd indices of x and subtracts 10 from the elements stored even indices of x. Calculate the standard deviation of this new vector.

Exercise 4

*# starter*

quiz\_list = list(

x = c(1, 2),

y = "Hello Quiz Taker",

z = "z"

)

Which of the following would return the third element of the list quiz\_list?

quiz\_list[3]

quiz\_list[[3]]

quiz\_list["3"]

quiz\_list$z

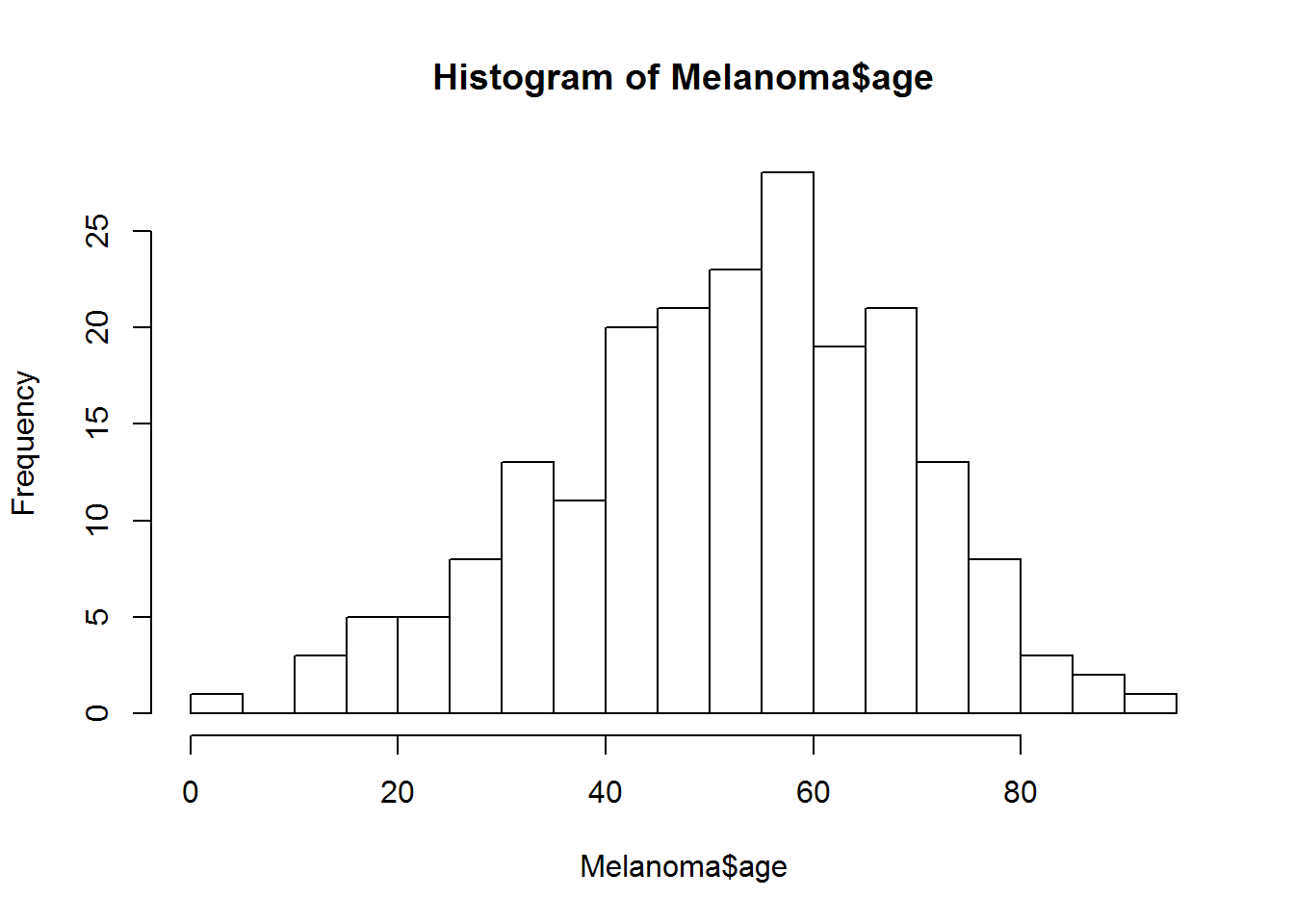
quiz\_list$3

Exercise 5

*# starter*

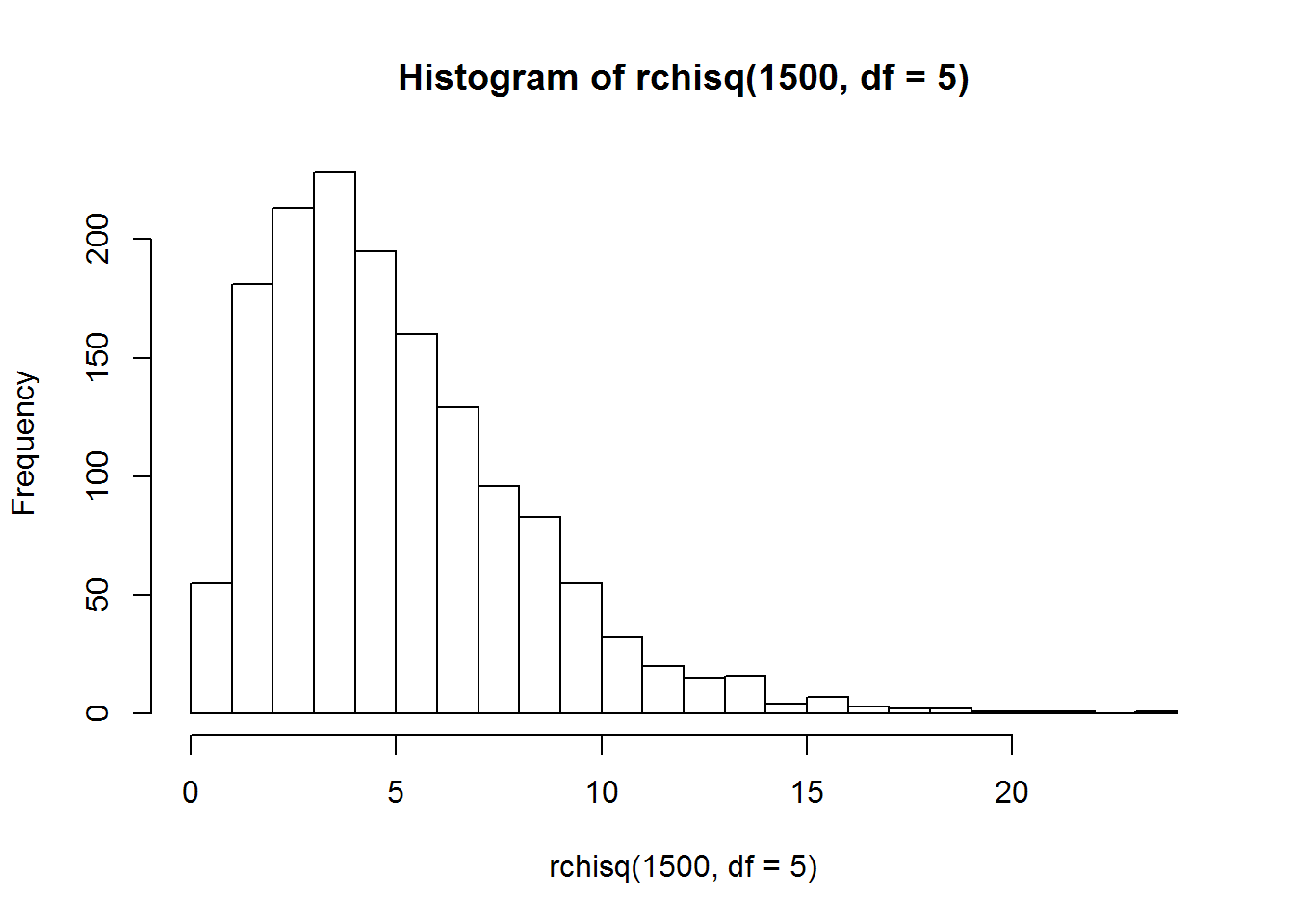
**library**(MASS)

Create a histogram of age in the Melanoma dataset from the MASS package. How would you describe this data?

* Large left skew
* **Slight left skew**
* Symmetric
* Slight right skew
* Large right skew

Possible hint: For an example of a right skew, run:

hist(rchisq(1500, df = 5), breaks = 20)

****Graded

Exercise 1

How many individuals in the Melanoma dataset from the MASS package died from a melanoma?

Exercise 2

What is the average age of individuals in the Melanoma dataset from the MASS package who are alive?

Exercise 3

Which animal in the mammals dataset from the MASS package has the largest brain weight relative to its body weight? (That is, the largest brain weight to body weight ratio.)

Exercise 4

Create side-by-side boxplots for each of the numeric variables in the iris dataset. To do so, simply supply the usual function with a dataframe of only the numeric variales of the dataset. Based on this plot, which variable is the most variable? Calculate the standard deviation of this variable.

Exercise 5

*# preamble*

set.seed(42)

z = list(

round(rnorm(n = 25, 0, 5), 2),

c(1, 1, 2, 3, 5, 8),

sample(30)

)

The above code block has access to a list stored in the variable z.

Calculate the sum of:

* The minimum first element of z
* The maximum of the second element of z
* The mean of the third element of z

Exercise 6

Where were the measurements taken in the airquality dataset?

* Chicago, IL
* Los Angeles, CA
* **New York, NY**
* Champaign, IL
* Paris, France

Exercise 7

Using the airquality dataset, what is the average wind speed in May ?

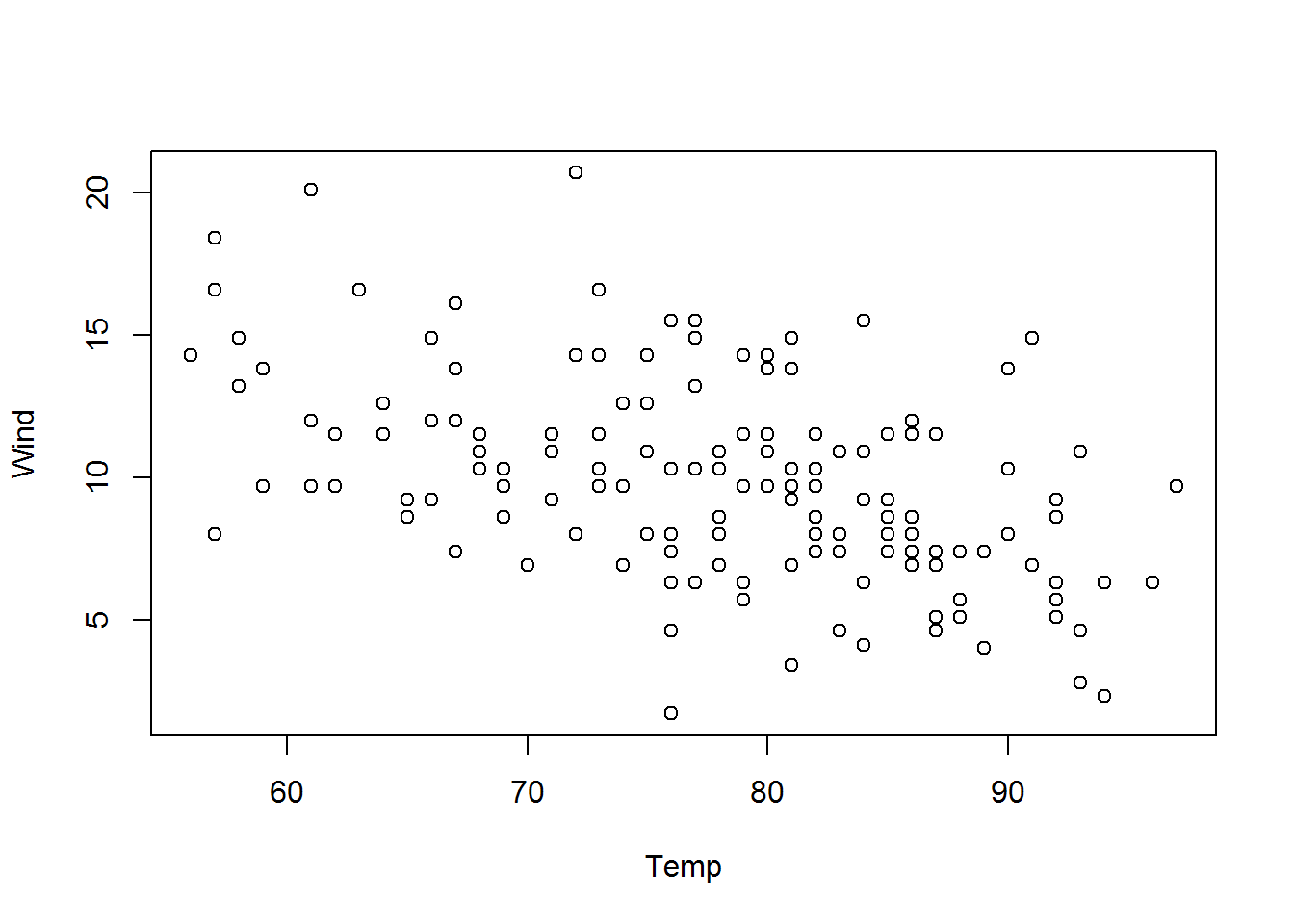
Exercise 8

Using the airquality dataset, what is the average ozone measurement? Hint: read the documentation of any function that returns an unexpected result. You will likely find a solution to the issue.

Exercise 9

Using the airquality dataset, create a scatter plot to compare windspeed and temperature. Based on this plot, you believe that:

* Wind speed and temperature have no relationship.
* As temperature increases, wind speed increases.
* **As temperature increases, wind speed decreases**

****Exercise 10

*# starter*

set.seed(1337)

x = rnorm(10000)

What proportion of the elements of x are larger than 2 in magnitude? Be sure to run the two lines in order, otherwise your vector will not contain the expected elements.

Exercise 11

*# starter*

set.seed(42)

x = rnorm(100, mean = 0, sd = 10)

mean(f(input = x)) - f()

Write a function called f that has a single argument input with a default value of 42 which is assumed to be a vector of numeric valves. The function should output a vector that is input but with any negative values replaced with 0.

Hint: The ifelse() function could be useful here. Note that all three arguments to ifelse() are vectors.

Run your function followed by the three lines given. Submit the output as your answer.

Exercise 12

*# starter*

set.seed(42)

y = 5 \* x0 + x1 + rnorm(n = 30, mean = 0 , sd = 1)

Create three vectors x0, x1, and y. Each should have a length of 30 and store the following:

x0: Each element should be the value 1 x1: The first 30 square numbers, starting from 1 (so 1, 4, 9, etc.) y: The result of running the given code, after creating the other two vectors

Report the mean of the values stored in y.

Exercise 13

**(Continued from Exercise 12)** Create a matrix X with columns x0 and x1. Report the sum of the elements in rows 17 and 19.

Exercise 14

**(Continued from Exercises 12 and 13)** Use matrix operations to create a new matrix beta\_hat defined as follows. Report the sum of the values stored in this matrix.

β^=(XTX)−1XTyβ^=(XTX)−1XTy

Exercise 15

**(Continued from Exercises 12, 13, and 14)** Create a new variable y\_hat which stores the result of the matrix operation,

y^=Xβ^.y^=Xβ^.

The result will be a 30×130×1 matrix. Perform and report the result of the following operation,

∑i=130(yi−y^i)2.∑i=130(yi−y^i)2.

Week 2 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 XX that has a normal distribution with a mean of 5 and a variance of 9. Calculate P[X>4]P[X>4].

Exercise 2

*#*

Consider the simple linear regression model

Y=−3+2.5x+ϵY=−3+2.5x+ϵ

where

ϵ∼N(0,σ2=4).ϵ∼N(0,σ2=4).

What is the expected value of YY given that x=5x=5? That is, what is E[Y∣X=5]E[Y∣X=5]?

Exercise 3

*#*

Given the SLR model in exercise 2, what is the standard deviation of YY when xx is 1010. That is, what is SD[Y∣X=10]SD[Y∣X=10]?

Exercise 4

*#*

For this Exercise, use the built-in trees dataset in R. Fit a simple linear regression model with Girth as the response and Height as the predictor. What is the slope of the fitted regression line?

Exercise 5

*#*

Continue using the SLR model you fit in Exercise 4. What is the value of R2R2 for this fitted SLR model?

Graded

Exercise 1

*#*

Consider the simple linear regression model

Y=10+5x+ϵY=10+5x+ϵ

where

ϵ∼N(0,σ2=16pσ2=16).

Calculate the probability that YY is less than 6 given that x=0x=0.

Exercise 2

*#*

Using the SLR model in exercise 1, what is the probability that YY is greater than 3 given that x=−1x=−1?

Exercise 3

*#*

Using the SLR model in exercise 1, what is the probability that YY is greater than 3 given that x=−2x=−2?

Exercise 4

*#*

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

Suppose we would like to predict the duration of an eruption of [the Old Faithful geyser](http://www.yellowstonepark.com/about-old-faithful/) in [Yellowstone National Park](https://en.wikipedia.org/wiki/Yellowstone_National_Park) based on the waiting time before an eruption. Fit a simple linear model in R that accomplishes this task.

What is the estimate of the intercept parameter?

Exercise 5

*#*

What is the estimate of the slope parameter?

Exercise 6

*#*

Use the fitted model to predict the duration of an eruption based on a waiting time of **80** minutes.

Exercise 7

*#*

Use the fitted model to predict the duration of an eruption based on a waiting time of **120** minutes.

Exercise 8

*#*

Of the predictions that you made, for 80 and 120 minutes, which is more reliable?

* **80**
* 120
* Both are equally reliable

Exercise 9

*#*

Calculate the RSS for the fitted model.

Exercise 10

*#*

What proportion of the variation in eruption duration is explained by the linear relationship with waiting time?

Exercise 11

*#*

Calculate the standard deviation of the residuals of the fitted model.

Exercise 12

Suppose both Least Squares and Maximum Likelihood are used to fit a simple linear regression model to the same data. The estimates for the slope and the intercept will be:

* **The same same**
* **Different**
* **Possibly the same or different depending on the data**

Exercise 13

Consider the fitted regression model:

y^=−1.5+2.3xy^=−1.5+2.3x

Indicate all of the following that **must** be true:

* **The difference between the yy values of observations at x=10x=10 and x=9x=9 is 2.32.3.**
* **A good estimate for the mean of YY when x=0x=0 is -1.5.**
* **There are observations in the dataset used to fit this regression with negative yy values.**

Exercise 14

Indicate all of the following that are true:

* The SLR model assumes that errors are independent. true
* The SLR model allows for a larger variances for larger values of the predictor variable.
* The SLR model assumes that the response variable follows a normal distribution.
* The SLR model assumes taht the relationship between the response and the predictor is linear. true

Exercise 15

Suppose you fit a simple linear regression model and obtain β^1=0β^1=0. Does this mean that there is **no relationship** between the response and the predictor?

* **Yes**
* **No no**
* **Depends on the intercept**

# Week 3 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 XX that has a tt distribution with 77 degrees of freedom. Calculate P[X>1.3]P[X>1.3].

### Exercise 2

*#*

Consider a random variable YY that has a tt distribution with 99 degrees of freedom. Find cc such that P[X>c]=0.025P[X>c]=0.025.

### Exercise 3

*#*

For this Exercise, use the built-in trees dataset in R. Fit a simple linear regression model with Girth as the response and Height as the predictor. What is the p-value for testing H0:β1=0H0:β1=0 vs H1:β1≠0H1:β1≠0?

### Exercise 4

*#*

Continue using the SLR model you fit in Exercise 3. What is the length of a 90% confidence interval for β1β1?

### Exercise 5

*#*

Continue using the SLR model you fit in Exercise 3. Calculate a 95% confidence interval for the mean tree girth of a tree that is 79 feet tall. Report the upper bound of this interval.

## Graded

### Exercise 1

*#*

Consider a random variable XX that has a tt distribution with 55 degrees of freedom. Calculate P[|X|>2.1]P[|X|>2.1].

### Exercise 2

*#*

Calculate the critical value used for a 90% confidence interval about the slope parameter of a simple linear regression model that is fit to 10 observations. (Your answer should be a positive value.)

### Exercise 3

*#*

Consider the true simple linear regression model

Yi=5+4xi+ϵiϵi∼N(0,σ2=4)i=1,2,…20Yi=5+4xi+ϵiϵi∼N(0,σ2=4)i=1,2,…20

Given Sxx=1.5Sxx=1.5, calculate the probability of observing data according to this model, fitting the SLR model, and obtaining an estimate of the slope parameter greater than 4.2. In other words, calculate

P[β^1>4.2]P[β^1>4.2]

### Exercise 4

*#*

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

Suppose we would like to predict the duration of an eruption of [the Old Faithful geyser](http://www.yellowstonepark.com/about-old-faithful/) in [Yellowstone National Park](https://en.wikipedia.org/wiki/Yellowstone_National_Park) based on the waiting time before an eruption. Fit a simple linear model in R that accomplishes this task.

What is the value of SE[β^1]SE[β^1]?

### Exercise 5

*#*

What is the value of the test statistic for testing H0:β0=0H0:β0=0 vs H1:β0≠0H1:β0≠0?

### Exercise 6

*#*

What is the value of the test statistic for testing H0:β1=0H0:β1=0 vs H1:β1≠0H1:β1≠0?

### Exercise 7

*#*

Test H0:β1=0H0:β1=0 vs H1:β1≠0H1:β1≠0 with α=0.01α=0.01. What decision do you make?

* **Fail to reject H0H0**
* **Reject H0H0 yes**
* **Reject H1H1**
* **Not enough information**

### Exercise 8

*#*

Calculate a 90% confidence interval for β0β0. Report the upper bound of this interval.

### Exercise 9

*#*

Calculate a 95% confidence interval for β1β1. Report the length of the margin of this interval.

### Exercise 10

*#*

Create a 90% confidence interval for the mean eruption duration for a waiting time of 81 minutes. Report the lower bound of this interval.

### Exercise 11

*#*

Create a 99% prediction interval for a new observation’s eruption duration for a waiting time of 72 minutes. Report the upper bound of this interval.

### Exercise 12

Consider a 90% confidence interval for the mean response and a 90% prediction interval, both at the same xx value. Which interval is narrower?

* **Confidence interval yes**
* **Prediction interval**
* **No enough information, it depends on the value of xx**

### Exercise 13

Suppose you obtain a 99% confidence interval for β1β1 that is (−0.4,5.2)(−0.4,5.2). Now test H0:β1=0H0:β1=0 vs H1:β1≠0H1:β1≠0 with α=0.01α=0.01. What decision do you make?

* **Fail to reject**H0H0 yes
* Reject H0H0
* Reject H1H1
* Not enough information

### Exercise 14

Suppose you test H0:β1=0H0:β1=0 vs H1:β1≠0H1:β1≠0 with α=0.01α=0.01 and fail to reject H0H0. Indicate all of the following that must always be true:

* **There is no relationship between the response and the predictor. no**
* **The probability of observing the estimated value of β1β1 (or something more extreme) is greater than 0.010.01 if we assume that β1=0β1=0. yes**
* **The value of β^1β^1 is very small. For example, it could not be 1.2. no**
* **The probability that β1=0β1=0 is very high. no**
* **We would also fail to reject at α=0.05α=0.05. no**

### Exercise 15

Consider a 95% confidence interval for the mean response calculated at x=6x=6. If instead we calculate the interval at x=7x=7, mark each value that would change:

* **Estimate yes**
* Critical Value
* **Standard Error yes**

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 XX that has a FF distribution with 33 and 55 degrees of freedom. Calculate P[X>2.7]P[X>2.7].

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=β0+β1x1+β2x2+β3x3+ϵY=β0+β1x1+β2x2+β3x3+ϵ

where

* x1x1 is GNP
* x2x2 is Population
* x3x3 is Armed.Forces

Create a 90% confidence interval for β1β1. Report the lower bound of this interval.

Exercise 3

*#*

What is the standard error of β^2β^2?

Exercise 4

*#*

What is the p-value for testing H0:β3=0 vs H1:β3≠0H0:β3=0 vs H1:β3≠0?

Exercise 5

*#*

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

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 FF test statistic is 2.4, what is the p-value of this test?

Exercise 2

*#*

What is the p-value for testing H0:β1=0 vs H1:β1≠0H0:β1=0 vs H1:β1≠0 in a multiple linear regression model with 5 predictors and 20 observations if the value of the tt test statistic is -1.3?

Exercise 3

set.seed(42)

x\_values = data.frame(

x1 = runif(15),

x2 = runif(15),

x3 = runif(15)

)

Consider the true model

Y=3+2x1+0.5x2+5x3+ϵY=3+2x1+0.5x2+5x3+ϵ

where

ϵ∼N(0,σ2=9)ϵ∼N(0,σ2=9)

What is SD[β^2]SD[β^2] given the values of predictors above?

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

Exercise 5

*#*

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

Exercise 6

*#*

Calculate the p-value of the test H0:βExamination=0 vs H1:βExamination≠0H0:βExamination=0 vs H1:βExamination≠0

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.

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.

Exercise 9

*#*

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

Exercise 10

*#*

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

* Fail to reject H0H0
* **Reject H0H0**
* Reject H1H1
* Not enough information

Exercise 11

*#*

Consider a model that only uses the predictors Education, Catholic, and Infant.Mortality. Use an FF test to compare this with the model that uses all predictors. Report the p-value of this test.

Exercise 12

Consider two **nested** multiple linear regression models fit to the same data. One has an R2R2 of 0.9 while the other has an R2R2 of 0.8. Which model uses fewer predictors?

* The model with an R2R2 of 0.9
* The model with an R2R2 of 0.8
* Not enough information

Exercise 13

The following multiple linear regression is fit to data

Y=β0+β1x1+β2x2+ϵY=β0+β1x1+β2x2+ϵ

If β^1=5β^1=5 and β^2=0.25β^2=0.25 then:

* The p-value for testing H0:β1=0 vs H1:β1≠0H0:β1=0 vs H1:β1≠0 will be *larger than* the p-value for testing H0:β2=0 vs H1:β2≠0H0:β2=0 vs H1:β2≠0.
* The p-value for testing H0:β1=0 vs H1:β1≠0H0:β1=0 vs H1:β1≠0 will be *smaller than* the p-value for testing H0:β2=0 vs H1:β2≠0H0:β2=0 vs H1:β2≠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=β0+β1x1+β2x2+ϵY=β0+β1x1+β2x2+ϵ

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

* The p-values for both H0:β1=0 vs H1:β1≠0H0:β1=0 vs H1:β1≠0 and H0:β2=0 vs H1:β2≠0H0:β2=0 vs H1:β2≠0 will be less than 0.01.
* **The p-values for both H0:β1=0 vs H1:β1≠0H0:β1=0 vs H1:β1≠0 and H0:β2=0 vs H1:β2≠0H0:β2=0 vs H1:β2≠0 could be greater than 0.01.**
* H0:β1=0 vs H1:β1≠0H0:β1=0 vs H1:β1≠0 will have a p-value less than 0.01 if H0:β2=0 vs H1:β2≠0H0:β2=0 vs H1:β2≠0 has a p-value greater than 0.01.

Week 7 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

*#*

For each of the following Exercises, use the built-in ToothGrowth dataset in R. We will use len as the response variable, which we will refer to as the tooth length. Use ?ToothGrowth to learn more about the dataset.

For Exercises 1-3, consider the dose variable a numeric variable. Fit the regression model

Y=β0+β1x1+β2x2+β3x1x2+ϵY=β0+β1x1+β2x2+β3x1x2+ϵ

where

* YY is tooth length
* x1x1 is the dose in milligrams per day
* x2x2 is a dummy variable that takes the value 11 when the supplement type is ascorbic acid

Use this model to obtain an estimate of the change in mean tooth length for an dose increase of 1 milligram per day, when the supplement type is orange juice.

Exercise 2

*#*

Use the model from Exercise 1 to obtain an estimate of the change in mean tooth length for an dose increase of 1 milligram per day, when the supplement type is ascorbic acid.

Exercise 3

*#*

The answers to the two previous questions should be different, but are these results significant? Test for interaction between dose and supplement type. Report the p-value of the test.

Exercise 4

*#*

unique(ToothGrowth$dose)

## [1] 0.5 1.0 2.0

Note that there are only three unique values for the dosages. For Exercises 4 and 5, consider the dose variable a categorical variable.

The previous model, using dose as numeric, assumed that the difference between a dose of 0.5 and 1.0 is the same as the difference between a dose of 1.0 and 1.5, but allowed us to make predictions for any dosage.

Considering dose a categorical variable, we will only be able to make predictions at the three existing dosages, but no longer is the the relationship between dose and response constrained to be linear.

Fit the regression model

Y=β0+β1x1+β2x2+β3x3+ϵY=β0+β1x1+β2x2+β3x3+ϵ

where

* YY is tooth length
* x1x1 is a dummy variable that takes the value 11 when the dose is 1.0 milligrams per day
* x2x2 is a dummy variable that takes the value 11 when the dose is 2.0 milligrams per day
* x3x3 is a dummy variable that takes the value 11 when the supplement type is ascorbic acid

Use this model to obtain an estimate of the difference in mean tooth length for dosages of 1.0 and 2.0 milligrams per day for both supplement types. (Since we are not considering interactions, the supplement type does not matter.)

Exercise 5

Suppose we wrote the the previous model with a different parameterization

Y=γ1x1+γ2x2+γ3x3+γ4x4+ϵY=γ1x1+γ2x2+γ3x3+γ4x4+ϵ

where

* YY is tooth length
* x1x1 is a dummy variable that takes the value 11 when the dose is 0.5 milligrams per day
* x2x2 is a dummy variable that takes the value 11 when the dose is 1.0 milligrams per day
* x3x3 is a dummy variable that takes the value 11 when the dose is 2.0 milligrams per day
* x4x4 is a dummy variable that takes the value 11 when the supplement type is ascorbic acid

Calculate an estimate of γ3γ3.

Graded

Exercise 1

*# starter*

**library**(MASS)

For exercises 1 - 6, use the cats dataset from the MASS package. Consider three models:

* Simple: Y=β0+β1x1+ϵY=β0+β1x1+ϵ
* Additive: Y=β0+β1x1+β2x2+ϵY=β0+β1x1+β2x2+ϵ
* Interaction: Y=β0+β1x1+β2x2+β3x1x2+ϵY=β0+β1x1+β2x2+β3x1x2+ϵ

where

* YY is the heart weight of a cat in grams
* x1x1 is the body weight of a cat in kilograms
* x2x2 is a dummy variable that takes the value 11 when a cat is male

Use the simple model to estimate the change in average heart weight when body weight is increased by 1 kilogram, for a female cat.

Exercise 2

*# starter*

**library**(MASS)

Use the interaction model to estimate the change in average heart weight when body weight is increased by 1 kilogram, for a female cat.

Exercise 3

*# starter*

**library**(MASS)

Use the interaction model to estimate the change in average heart weight when body weight is increased by 1 kilogram, for a male cat.

Exercise 4

*# starter*

**library**(MASS)

Use the additive model to estimate the **difference** in the change in average heart weight when body weight is increased by 1 kilogram between a male and female cats.

Exercise 5

*# starter*

**library**(MASS)

Use an FF test to compare the additive and interaction models. Report the value of the FF test statistic.

Exercise 6

*# starter*

**library**(MASS)

Carry out the test in Exercise 5 using α=0.05α=0.05. Based on this test, which model is preferred?

* Simple
* Additive
* Interaction
* None of the above

Exercise 7

*# starter*

iris\_add = lm(Sepal.Length ~ Petal.Length + Species, data = iris)

For exercises 7 - 13, use the iris dataset which is built into R. Use ?iris to learn about this dataset. (Note that this model would be somewhat odd in practice. Usually it would make sense to predict species from characteristics, or characteristics from species. Here we’re using a combination of characteristics and species to predict other characteristics, for illustrative purposes.)

Using the model fit with the given code, predict the sepal length of a versicolor with a petal length of 5.10.

Exercise 8

*# starter*

iris\_add = lm(Sepal.Length ~ Petal.Length + Species, data = iris)

Continue to use the model from Exercise 7. Create a 90% confidence interval for the difference in mean sepal length between virginicas and setosas for a given petal length. Report the lower bound of this interval

Exercise 9

*# starter*

iris\_add = lm(Sepal.Length ~ Petal.Length + Species, data = iris)

Continue to use the model from Exercise 7. Perform a test that compares this model to one without an effect for species. Report the value of the test statistic for this test.

Exercise 10

*# starter*

iris\_int = lm(Sepal.Length ~ Petal.Length \* Species, data = iris)

Now consider the model with interaction given above. Excluding σ2σ2, how many parameters does this model have? Stated another way, if written mathematically, how many ββ parameters are in the model?

Exercise 11

*# starter*

iris\_int = lm(Sepal.Length ~ Petal.Length \* Species, data = iris)

Using the interaction model fit with the given code, create a 99% prediction interval for the sepal length of a versicolor with a petal length of 5.10. Report the upper bound of this interval.

Exercise 12

*# starter*

iris\_int = lm(Sepal.Length ~ Petal.Length \* Species, data = iris)

Using the interaction model fit with the given code, obtain an estimate of the change in mean petal length for a sepal length increase of 1 unit, for a versicolor.

Exercise 13

*# starter*

Compare the two previous models, the additive and interaction models using an ANVOA FF test using α=0.01α=0.01. Based on this test, which model is preferred?

* Additive
* Interaction
* None of the above

Exercise 14

*# starter*

For exercises 14 - 15, use the swiss dataset, which is built into R. Fit an multiple linear model with Fertility as the response and Education, Catholic, and Infant.Mortality as predictors. Use the first order terms as well as all two and three-way interactions.

Use this model to estimate the change in mean Fertility for an increase of Education of one unit when Catholic is 90.0 and Infant.Mortality is 20.0.

Exercise 15

*# starter*

Test for the significance of the three-way interaction in model from Exercise 14. Report the p-value of this test.

Week 8 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

*# starter*

Consider the model

Y=5−2x+ϵY=5−2x+ϵ

where

ϵ∼N(0,σ2=|x|4).ϵ∼N(0,σ2=|x|4).

That is

Var[Y∣X=x]=|x|4.Var[Y∣X=x]=|x|4.

Calculate

P[Y>1∣X=3].P[Y>1∣X=3].

Exercise 2

*# preamble*

gen\_data = **function**(sample\_size = 20, seed = 420) {

set.seed(seed)

x = runif(n = sample\_size, min = 0, max = 3)

y = exp(2 + 3 \* x + 0.35 \* x ^ 2 + rnorm(n = sample\_size, sd = 3))

data.frame(x = x, y = y)

}

quiz\_data = gen\_data()

*# starter*

quiz\_data

The above code block has access to a data frame named quiz\_data with two variables y and x. Here, we use y as the response.

Fit a simple linear regression model to this data. What is the Cook’s distance for the observation with the largest leverage?

Exercise 3

*# preamble*

gen\_data = **function**(sample\_size = 20, seed = 420) {

set.seed(seed)

x = runif(n = sample\_size, min = 0, max = 3)

y = exp(2 + 3 \* x + 0.35 \* x ^ 2 + rnorm(n = sample\_size, sd = 3))

data.frame(x = x, y = y)

}

quiz\_data = gen\_data()

*# starter*

quiz\_data

The above code block has access to a data frame named quiz\_data with two variables y and x. Here, we use y as the response.

Fit a simple linear regression model to this data. Calculate the p-value of the Shapiro-Wilk test for the normality assumption.

Exercise 4

*# preamble*

gen\_data = **function**(sample\_size = 20, seed = 420) {

set.seed(seed)

x = runif(n = sample\_size, min = 0, max = 3)

y = exp(2 + 3 \* x + 0.35 \* x ^ 2 + rnorm(n = sample\_size, sd = 3))

data.frame(x = x, y = y)

}

quiz\_data = gen\_data()

*# starter*

quiz\_data

The above code block has access to a data frame named quiz\_data with two variables y and x. Here, we use y as the response.

Fit the model

log(y)=β0+β1x+β2x2+ϵ.log⁡(y)=β0+β1x+β2x2+ϵ.

Use the Shapiro-Wilk test to assess the normality assumption for this model. Use α=0.05α=0.05.

Select the correct decision and interpretation:

* Fail to Reject H0H0. Normality assumption is suspect.
* Fail to Reject H0H0. Normality assumption is *not* suspect.
* Reject H0H0. Normality assumption is suspect.
* Reject H0H0. Normality assumption is *not* suspect.
* Hint: The null hypothesis of the test assumes normality.

Exercise 5

*# preamble*

gen\_data = **function**(sample\_size = 20, seed = 420) {

set.seed(seed)

x = runif(n = sample\_size, min = 0, max = 3)

y = exp(2 + 3 \* x + 0.35 \* x ^ 2 + rnorm(n = sample\_size, sd = 3))

data.frame(x = x, y = y)

}

quiz\_data = gen\_data()

*# starter*

quiz\_data

The above code block has access to a data frame named quiz\_data with two variables y and x. Here, we use y as the response.

Fit the model

log(y)=β0+β1x+β2x2+ϵ.log⁡(y)=β0+β1x+β2x2+ϵ.

Calculate the residual sum of squares (RSS) in the original units of yy. That is, calculate

∑(y^i−yi)2.∑(y^i−yi)2.

Report your answer in billions.

Graded

Exercise 1

*# preamble*

gen\_data\_1 = **function**(sample\_size = 25, seed = 420) {

set.seed(seed)

x = runif(n = sample\_size, min = 0, max = 3)

y = 2 + 3 \* x + rnorm(n = sample\_size)

data.frame(x = x, y = y)

}

gen\_data\_2 = **function**(sample\_size = 25, seed = 420) {

set.seed(seed)

x = runif(n = sample\_size, min = 0, max = 3)

y = 2 + 3 \* x + rt(n = sample\_size, df = 2)

data.frame(x = x, y = y)

}

data\_1 = gen\_data\_1()

data\_2 = gen\_data\_2()

*# starter*

data\_1

data\_2

The above code block has access to two data frames named data\_1 and data\_2, both with variables y and x. Here, we use y as the response.

Fit a simple linear regression to both datasets. For both fitted regressions, create a Normal Q-Q Plot.

Based on the plots:

* The normality assumption is more suspect for the model fit to data\_1.
* The normality assumption is more suspect for the model fit to data\_2.

Exercise 2

*# preamble*

gen\_data\_2 = **function**(sample\_size = 100, seed = 420) {

set.seed(seed)

x = runif(n = sample\_size, min = 0, max = 3)

y = 2 + 3 \* x + rnorm(n = sample\_size)

data.frame(x = x, y = y)

}

gen\_data\_1 = **function**(sample\_size = 100, seed = 420) {

set.seed(seed)

x = runif(n = sample\_size, min = -3, max = 0)

y = 2 + 3 \* x + sqrt(abs(x \* rnorm(n = sample\_size)))

data.frame(x = x, y = y)

}

data\_1 = gen\_data\_1()

data\_2 = gen\_data\_2()

*# starter*

data\_1

data\_2

The above code block has access to two data frames named data\_1 and data\_2, both with variables variables y and x. Here, we use y as the response.

Fit a simple linear regression to both datasets. For both fitted regressions, create Fitted versus Residuals plot.

Based on the plots:

* The equal variance assumption is more suspect for the model fit to data\_1.
* The equal variance assumption is more suspect for the model fit to data\_2.

Exercise 3

*# starter*

Consider the model

Y=2+4x+ϵY=2+4x+ϵ

where

ϵ∼N(0,σ2=x2).ϵ∼N(0,σ2=x2).

That is

Var[Y∣X=x]=x2.Var[Y∣X=x]=x2.

Calculate

P[Y<−12∣X=−3].P[Y<−12∣X=−3].

Exercise 4

*# starter*

For exercises 4 - 9, use the LifeCycleSavings dataset which is built into R. Fit a multiple linear regression model with sr as the response and the remaining variables as predictors. What proportion of observations have a standardized residual less than 2 in magnitude?

Exercise 5

*# starter*

Continue using the model fit in Exercise 4. Note that each observation is about a particular country. Which country (observation) has the standardized residual with the largest magnitude?

Exercise 6

*# starter*

Continue using the model fit in Exercise 4. How many observations have “high” leverage? Use twice the average leverage as the cutoff for “high.”

Exercise 7

*# starter*

Continue using the model fit in Exercise 4. Which country (observation) has the largest leverage?

Exercise 8

*# starter*

Continue using the model fit in Exercise 4. Report the largest Cook’s Distance for observations in this dataset.

Exercise 9

*# starter*

Continue using the model fit in Exercise 4. Find the observations that are influential. Use 4n4n as the cutoff for labeling an observation influential.

Create a subset of the original data that excludes these influential observations and refit the same model to this new data. Report the sum of the estimated regression coefficients.

Exercise 10

*# starter*

airquality = na.omit(airquality)

For exercises 10 - 15, use the airquality dataset which is built into R. For simplicity, we will remove any observations with missing data. We will use Ozone as the response and Temp as a single predictor.

Fit the model

Y=β0+β1x+β2x2+ϵY=β0+β1x+β2x2+ϵ

Test for the significance of the quadratic term. Report the p-value of this test.

Exercise 11

*# starter*

airquality = na.omit(airquality)

Fit the model

Y=β0+β1x+β2x2+β3x3+β4x4+ϵY=β0+β1x+β2x2+β3x3+β4x4+ϵ

Test to compare this model to the model fit in Exercise 10. Report the p-value of this test.

Exercise 12

*# starter*

airquality = na.omit(airquality)

Use the Shapiro-Wilk test to asses the normality assumption for the model in Exercise 11. Use α=0.01α=0.01.

Select the correct decision and interpretation:

* Fail to Reject H0H0. Normality assumption is suspect.
* Fail to Reject H0H0. Normality assumption is *not* suspect.
* Reject H0H0. Normality assumption is suspect.
* Reject H0H0. Normality assumption is *not* suspect.

Exercise 13

*# starter*

airquality = na.omit(airquality)

Fit the model

log(y)=β0+β1x+ϵ.log⁡(y)=β0+β1x+ϵ.

Use the Shapiro-Wilk test to asses the normality assumption for this model. Use α=0.01α=0.01.

Select the correct decision and interpretation:

* Fail to Reject H0H0. Normality assumption is suspect.
* Fail to Reject H0H0. Normality assumption is *not* suspect.
* Reject H0H0. Normality assumption is suspect.
* Reject H0H0. Normality assumption is *not* suspect.

Exercise 14

*# starter*

airquality = na.omit(airquality)

Use the model from Exercise 13 to create a 90% prediction interval for Ozone when the temperate is 84 degree Fahrenheit. Report the upper bound of this interval

Exercise 15

*# starter*

airquality = na.omit(airquality)

Using the model from Exercise 13, calculate the ratio of:

* The sample variance of residuals for observations with a fitted value less than 3.5
* The sample variance of residuals for observations with a fitted value greater than 3.5

(While not a formal test for the equal variance assumption, we would hope that this value is close to 1.)

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)

}

*# 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.

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.

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.

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 R2R2 of the model with the better Adjusted R2R2.

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.

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.)

Exercise 2

*# starter*

mtcars

What is the Adjusted R2R2 of the model fit in Exercise 1?

Exercise 3

*# starter*

mtcars

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

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.)

* 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?

Exercise 6

*# starter*

mtcars

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

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.)

* 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?

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.

Exercise 11

*# starter*

LifeCycleSavings

Fit a model with all available predictors as well as their two-way interactions. What is the Adjusted R2R2 of this model?

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.)

* 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.)

* 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.

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 R2R2, which of these models is best? Report the Adjusted R2R2 of the model you choose.

Week 10 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*

*# starter*

Consider a categorical response YY which takes possible values 00 and 11 as well as two numerical predictors X1X1 and X2X2. Recall that

p(x)=P[Y=1∣X=x]p(x)=P[Y=1∣X=x]

Consider the model

log(p(x)1−p(x))=β0+β1x1+β2x2log⁡(p(x)1−p(x))=β0+β1x1+β2x2

together with parameters

* β0=2β0=2
* β1=−1β1=−1
* β2=−1β2=−1

Calculate P[Y=1∣X1=1,X2=0]P[Y=1∣X1=1,X2=0].

Exercise 2

*# preamble*

make\_sim\_data = **function**(n = 100) {

x1 = rnorm(n = n)

x2 = rnorm(n = n, sd = 2)

x3 = rnorm(n = n, sd = 3)

x4 = rnorm(n = n)

x5 = rnorm(n = n)

x6 = rnorm(n = n)

x7 = rnorm(n = n)

eta = -1 + 0.75 \* x2 + 2.5 \* x6

p = 1 / (1 + exp(-eta))

y = rbinom(n = n, 1, prob = p)

data.frame(y, x1, x2, x3, x4, x5, x6, x7)

}

set.seed(1)

quiz\_data = make\_sim\_data()

*# starter*

quiz\_data

Recall that

p(x)=P[Y=1∣X=x]p(x)=P[Y=1∣X=x]

Use the data available in the above code chunk stored in quiz\_data to fit the model

log(p(x)1−p(x))=β0+β1x1+β2x2+β3x3+β4x4+β5x5+β6x6+β7x7.log⁡(p(x)1−p(x))=β0+β1x1+β2x2+β3x3+β4x4+β5x5+β6x6+β7x7.

Report the value of the estimate for β2β2.

Exercise 3

*# preamble*

make\_sim\_data = **function**(n = 100) {

x1 = rnorm(n = n)

x2 = rnorm(n = n, sd = 2)

x3 = rnorm(n = n, sd = 3)

x4 = rnorm(n = n)

x5 = rnorm(n = n)

x6 = rnorm(n = n)

x7 = rnorm(n = n)

eta = -1 + 0.75 \* x2 + 2.5 \* x6

p = 1 / (1 + exp(-eta))

y = rbinom(n = n, 1, prob = p)

data.frame(y, x1, x2, x3, x4, x5, x6, x7)

}

set.seed(1)

quiz\_data = make\_sim\_data()

*# starter*

quiz\_data

Use the data available in the above code chunk stored in quiz\_data to fit the model

log(p(x)1−p(x))=β0+β1x1+β2x2+β3x3+β4x4+β5x5+β6x6+β7x7.log⁡(p(x)1−p(x))=β0+β1x1+β2x2+β3x3+β4x4+β5x5+β6x6+β7x7.

Use a Wald test to test H0:β3=0H0:β3=0 versus H1:β3≠0H1:β3≠0. Report the p-value of this test.

Exercise 4

*# preamble*

make\_sim\_data = **function**(n = 100) {

x1 = rnorm(n = n)

x2 = rnorm(n = n, sd = 2)

x3 = rnorm(n = n, sd = 3)

x4 = rnorm(n = n)

x5 = rnorm(n = n)

x6 = rnorm(n = n)

x7 = rnorm(n = n)

eta = -1 + 0.75 \* x2 + 2.5 \* x6

p = 1 / (1 + exp(-eta))

y = rbinom(n = n, 1, prob = p)

data.frame(y, x1, x2, x3, x4, x5, x6, x7)

}

set.seed(1)

quiz\_data = make\_sim\_data()

*# starter*

quiz\_data

Use the data available in the above code chunk stored in quiz\_data to fit the model

log(p(x)1−p(x))=β0+β1x1+β2x2+β3x3+β4x4+β5x5+β6x6+β7x7.log⁡(p(x)1−p(x))=β0+β1x1+β2x2+β3x3+β4x4+β5x5+β6x6+β7x7.

Using this as an initial model, use BIC and a backwards stepwise procedure to select a reduced model. Use likelihood ratio test to compare the initial model and the selected model. Report the p-value of this test.

Exercise 5

*# preamble*

make\_sim\_data = **function**(n = 100) {

x1 = rnorm(n = n)

x2 = rnorm(n = n, sd = 2)

x3 = rnorm(n = n, sd = 3)

x4 = rnorm(n = n)

x5 = rnorm(n = n)

x6 = rnorm(n = n)

x7 = rnorm(n = n)

eta = -1 + 0.75 \* x2 + 2.5 \* x6

p = 1 / (1 + exp(-eta))

y = rbinom(n = n, 1, prob = p)

data.frame(y, x1, x2, x3, x4, x5, x6, x7)

}

set.seed(1)

quiz\_data = make\_sim\_data()

*# starter*

quiz\_data

*# fit the model here*

set.seed(1)

*# calculate the metric here*

Use the data available in the above code chunk stored in quiz\_data to fit the model

log(p(x)1−p(x))=β0+β1x1+β2x2+β3x3+β4x4+β5x5+β6x6+β7x7.log⁡(p(x)1−p(x))=β0+β1x1+β2x2+β3x3+β4x4+β5x5+β6x6+β7x7.

Calculate the 5-fold cross-validation misclassification rate when using this model as a classifier that seeks to minimize the misclassification rate. Since the data will be split randomly, use the seed provided after fitting the model. Also, use the relevant function from the boot package to ensure your calculation uses the same splits for grading purposes. (Even with the same seed, the splits could be done differently.)

Graded

Exercise 1

*# preamble*

*# starter*

Consider a categorical response YY which takes possible values 00 and 11 as well as three numerical predictors X1X1, X2X2, and X3X3. Recall that

p(x)=P[Y=1∣X=x]p(x)=P[Y=1∣X=x]

Consider the model

log(p(x)1−p(x))=β0+β1x1+β2x2+β3x3log⁡(p(x)1−p(x))=β0+β1x1+β2x2+β3x3

together with parameters

* β0=−3β0=−3
* β1=1β1=1
* β2=2β2=2
* β2=3β2=3

Calculate P[Y=0∣X1=−1,X2=0.5,X2=0.25]P[Y=0∣X1=−1,X2=0.5,X2=0.25].

Exercise 2

*# preamble*

*# starter*

For Exercises 2 - 7, use the built-in R dataset mtcars. We will use this dataset to attempt to predict whether or not a car has a manual transmission.

Recall that

p(x)=P[Y=1∣X=x]p(x)=P[Y=1∣X=x]

Fit the model

log(p(x)1−p(x))=β0+β1x1+β2x2+β3x3log⁡(p(x)1−p(x))=β0+β1x1+β2x2+β3x3

where

* YY is am
* x1x1 is mpg
* x2x2 is hp
* x3x3 is qsec

Report the value of the estimate for β3β3.

Exercise 3

*# preamble*

*# starter*

Using the model fit in Exercise 2, estimate the change in log-odds that a car has a manual transmission for an increase in fuel efficiency of one mile per gallon.

Exercise 4

*# preamble*

*# starter*

Using the model fit in Exercise 2, estimate the log-odds that a car has a manual transmission for a car with a fuel efficiency of 19 miles per gallon, 150 horsepower, and a quarter mile time of 19 seconds.

Exercise 5

*# preamble*

*# starter*

Using the model fit in Exercise 2, estimate the probability that a car with a fuel efficiency of 22 miles per gallon, 123 horsepower, and a quarter mile time of 18 seconds has a manual transmission.

Exercise 6

*# preamble*

*# starter*

Use a likeliood ratio test to test

H0:β1=β2=β3=0H0:β1=β2=β3=0

for the model fit in Exercise 2. Report the test statistic of this test.

Exercise 7

*# preamble*

*# starter*

Recall that

p(x)=P[Y=1∣X=x]p(x)=P[Y=1∣X=x]

Fit the model

log(p(x)1−p(x))=β0+β1x1+β2x2+β3x3log⁡(p(x)1−p(x))=β0+β1x1+β2x2+β3x3

where

* YY is am
* x1x1 is mpg
* x2x2 is hp
* x3x3 is qsec

Use a Wald test to test H0:β2=0H0:β2=0 versus H1:β2≠0H1:β2≠0. Report the p-value of this test.

Exercise 8

*# preamble*

*# starter*

**library**(MASS)

For Exercises 8 - 15, we will use two related diabetes datasets about the [Pima Native Americans](https://en.wikipedia.org/wiki/Pima_people) from the MASS package; Pima.tr and Pima.te. For details, use ?MASS::Pima.tr. They are essentially a train (Pima.tr) and test (Pima.te) dataset that are pre-split.

Recall that

p(x)=P[Y=1∣X=x]p(x)=P[Y=1∣X=x]

Use to training data to fit the model

log(p(x)1−p(x))=β0+β1x1+β2x2+β3x21+β4x22+β5x1x2log⁡(p(x)1−p(x))=β0+β1x1+β2x2+β3x12+β4x22+β5x1x2

where

* YY is a binary categorical variable that takes the value 11 when an individual is diabetic according to WHO criteria, 00 if not
* x1x1 is glu
* x2x2 is ped

Report the estimate of β4β4.

Hint: You do not need to create a response variable with values 11 and 00, instead you can use the factor variable type.

Exercise 9

*# preamble*

*# starter*

**library**(MASS)

Use the model fit in Exercise 8 to obtain a predicted probability of diabetes for each of the individuals in the test dataset (Pima.te). What proportion of these probabilities are larger than 0.80?

Exercise 10

*# preamble*

*# starter*

**library**(MASS)

Fit an additive logistic regression to model the probability of diabetes using the train dataset, Pima.tr, which uses all available predictors in the dataset. Using this as an initial model, use AIC and a backwards stepwise procedure to select a reduced model. How many predictors are used in this reduced model?

Exercise 11

*# preamble*

*# starter*

**library**(MASS)

Fit a logistic regression to model the probability of diabetes using the train dataset, Pima.tr, which uses all available predictors in the dataset as well as all possible two-way interactions. Using this as an initial model, use AIC and a backwards stepwise procedure to select a reduced model. What is the deviance of this reduced model?

Exercise 12

*# preamble*

*# starter*

**library**(MASS)

**library**(boot)

*# fit the models here*

set.seed(42)

*# get cross-validated results for the polynomial model here*

set.seed(42)

*# get cross-validated results for the additive model here*

set.seed(42)

*# get cross-validated results for the model selected from additive model here*

set.seed(42)

*# get cross-validated results for the interaction model here*

set.seed(42)

*# get cross-validated results for the model selected from interaction model here*

Obtain 5-fold cross-validated misclassification rates for each of the previous 5 models used as classifiers that seek to minimize the misclassification rate. (The models from Exercises 8, 10, and 11) Since the data will be split randomly, use the seeds provided to obtain the cross-validated results after fitting the models. Also, use the relevant cross-validation function from the boot package to ensure your calculation uses the same splits for grading purposes. (Even with the same seed, the splits could be done differently.)

Report the best cross-validated misclassification rate of these five.

Exercise 13

*# preamble*

*# starter*

**library**(MASS)

Using the additive model previously fit to the training dataset, create a classifier that seeks to minimize the misclassification rate. Report the misclassification rate or this classifier in the test dataest.

Exercise 14

*# preamble*

*# starter*

**library**(MASS)

Using the additive model previously fit to the training dataset, create a classifier that seeks to minimize the misclassification rate. Report the sensitivity of this classifier in the test dataset.

Exercise 15

*# preamble*

*# starter*

**library**(MASS)

Using the additive model previously fit to the training dataset, create a classifier that classifies an individual as diabetic if their predicted probability of diabetes is greater than 0.30.3. Report the sensitivity of this classifier in the test dataset.