Week 4 - Homework

STAT 420, Summer 2017, Dalpiaz

Exercise 1 (Using `lm`)

For this exercise we will use the data stored in `nutrition.csv`. It contains the nutritional values per serving size for a large variety of foods as calculated by the USDA. It is a cleaned version totaling 5,138 observations and is current as of September 2015.

The variables in the dataset are:

- `ID`
- `Desc` - Short description of food
- `Water` - in grams
- `Calories`
- `Protein` - in grams
- `Fat` - in grams
- `Carbs` - Carbohydrates, in grams
- `Fiber` - in grams
- `Sugar` - in grams
- `Calcium` - in milligrams
- `Potassium` - in milligrams
- `Sodium` - in milligrams
- `VitaminC` - Vitamin C, in milligrams
- `Chol` - Cholesterol, in milligrams
- `Portion` - Description of standard serving size used in analysis

(a) Fit the following multiple linear regression model in `R`. Use `Calories` as the response and `Carbs`, `Fat`, and `Protein` as predictors.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i.$$

Here,

- Y_i is `Calories`.
- x_{i1} is `Carbs`.
- x_{i2} is `Fat`.
- x_{i3} is `Protein`.

Use an F-test to test the significance of the regression. Report the following:

- The null and alternative hypotheses
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$
- A conclusion in the context of the problem

When reporting these, you should explicitly state them in your document, not assume that a reader will find and interpret them from a large block of `R` output.

Solution:

```
nutrition = read.csv("nutrition.csv")
library(broom)
```

```
nut_cfp = lm(Calories ~ Carbs + Fat + Protein, data = nutrition)
summary(nut_cfp)
```

```
##
## Call:
## lm(formula = Calories ~ Carbs + Fat + Protein, data = nutrition)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -381.13   -3.46   -0.31    5.33  259.23
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.768066   0.492386   7.653 2.34e-14 ***
## Carbs       3.773605   0.009698 389.093  < 2e-16 ***
## Fat         8.804109   0.015305 575.248  < 2e-16 ***
## Protein     3.967269   0.026284 150.938  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.89 on 5134 degrees of freedom
## Multiple R-squared:  0.9889, Adjusted R-squared:  0.9889
## F-statistic: 1.524e+05 on 3 and 5134 DF,  p-value: < 2.2e-16
```

- $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- H_1 : At least one of $\beta_j \neq 0, j = 1, 2, 3$
- Test statistic: $F = 1.5244482 \times 10^5$
- P-value: 0. (although, not actually 0, but very small)
- Decision: **Reject** H_0 at $\alpha = 0.01$.
- Conclusion: There is a linear relationship between Calories and at least some of carbs, fat and protein.

Note that we used the `broom` package to obtain some results directly. This is a useful package for "cleaning" some of the default `R` output.

(b) Output only the estimated regression coefficients. Interpret all $\hat{\beta}_j$ coefficients in the context of the problem.

Solution:

```
coef(nut_cfp)
```

```
## (Intercept)       Carbs         Fat     Protein
##    3.768066    3.773605    8.804109    3.967269
```

- $\hat{\beta}_0 = 3.768066$ is the estimated Calories of a food with 0g carbs, 0g fat, and 0g protein.
- $\hat{\beta}_1 = 3.7736045$ is the estimated change in mean Calories for an increase of 1g carbs for a food with a certain fat and protein content.
- $\hat{\beta}_2 = 8.8041089$ is the estimated change in mean Calories for an increase of 1g fat for a food with a certain carb and protein content.
- $\hat{\beta}_3 = 3.9672687$ is the estimated change in mean Calories for an increase of 1g protein for a food with a certain carb and fat content.

(c) Use your model to predict the number of `Calories` in a Big Mac. According to McDonald's publicized nutrition facts, the Big Mac contains 47g of carbohydrates, 28g of fat, and 25g of protein.

Solution:

```
big_mac = data.frame(Carbs = 47, Fat = 28, Protein = 25)
predict(nut_cfp, newdata = big_mac)
```

```
##        1
## 526.8242
```

(d) Calculate the standard deviation, s_y, for the observed values in the Calories variable. Report the value of s_e from your multiple regression model. Interpret both estimates in the context of this problem.

Solution:

```
(s_y = sd(nutrition$Calories))
```

```
## [1] 179.2444
```

```
(s_e = glance(nut_cfp)$sigma)
```

```
## [1] 18.89119
```

- $s_y = 179.244356$ gives us an estimate of the variability of Calories, specifically how the observed Calorie data varies about its mean. (We could think of this as an estimate for how the observations vary in the model $Y_i = \beta_0 + \epsilon_i$.) This estimate does not use the predictors in any way.
- $s_e = 18.8911855$ gives us an estimate of the variability of the residuals of the model, specifically how the observed Calorie data varies about the fitted regression. This estimate does take into account the predictors.

(e) Report the value of R^2 for the model. Interpret its meaning in the context of the problem.

Solution:

```
glance(nut_cfp)$r.squared
```

```
## [1] 0.9888987
```

$R^2 = 0.9889$ tells us that 98.89% of the observed variation in Calories is explained by a linear relationship with carbs, fat, and protein.

(f) Calculate a 90% confidence interval for β_2. Give an interpretation of the interval in the context of the problem.

Solution:

```
confint(nut_cfp, level = 0.90)
```

```
##                   5 %     95 %
## (Intercept) 2.958016 4.578116
## Carbs       3.757649 3.789560
## Fat         8.778930 8.829288
## Protein     3.924027 4.010510
```

```
confint(nut_cfp, level = 0.90)[3,]
```

```
##      5 %     95 %
## 8.778930 8.829288
```

We are 90% confident that the true change in mean Calories for an increase of 1g fat is in this interval for particular values of protein and carbs.

(g) Calculate a 95% confidence interval for β_0. Give an interpretation of the interval in the context of the problem.

Solution:

```
confint(nut_cfp, level = 0.95)[1,]
```

```
##    2.5 %   97.5 %
## 2.802779 4.733353
```

We are 95% confident that the true mean Calories for a food with 0g of carbs, fat, and protein is in this interval. (This **is** possible. We are, for example, not measuring alcohol content, which does provide Calories.)

(h) Use a 99% confidence interval to estimate the mean Calorie content of a small order of McDonald's french fries that has 30g of carbohydrates, 11g of fat, and 2g of protein. Interpret the interval in context.

Solution:

```
small_fries = data.frame(Carbs = 30, Fat = 11, Protein = 2)
predict(nut_cfp, newdata = small_fries, interval = "confidence", level = 0.99)
```

```
##        fit      lwr      upr
## 1 221.7559 220.8924 222.6195
```

We are 99% confident that the true mean Calories for foods with 30g of carbohydrates, 11g of fat, and 2g of protein is in this interval.

(i) Use a 90% prediction interval to predict the Calorie content of new healthy menu item that has 11g of carbohydrates, 1.5g of fat, and 1g of protein. Interpret the interval in context.

Solution:

```
healthy_item = data.frame(Carbs = 11, Fat = 1.5, Protein = 1)
predict(nut_cfp, newdata = healthy_item, interval = "prediction", level = 0.90)
```

```
##        fit     lwr      upr
## 1 62.45115 31.3649 93.53739
```

We are 90% confident that the Calories for a food item with 11g of carbohydrates, 1.5g of fat, and 1g of protein is in this interval.

Exercise 2 (More `lm`)

For this exercise we will use the data stored in `goalies_cleaned.csv`. It contains career data for 462 players in the National Hockey League who played goaltender at some point up to and including the 2014-2015 season. The variables in the dataset are:

- `W` - Wins
- `GA` - Goals Against
- `SA` - Shots Against
- `SV` - Saves
- `SV_PCT` - Save Percentage
- `GAA` - Goals Against Average
- `SO` - Shutouts
- `MIN` - Minutes
- `PIM` - Penalties in Minutes

For this exercise we will consider three models, each with Wins as the response. The predictors for these models are:

- Model 1: Goals Against, Shots Against, Saves
- Model 2: Goals Against, Shots Against, Saves, Minutes, Penalties in Minutes
- Model 3: All Available

```r
# read in data
goalies_cleaned = read.csv("goalies_cleaned.csv")

# fit requested models
m1 = lm(W ~ GA + SA + SV, data = goalies_cleaned)
m2 = lm(W ~ GA + SA + SV + MIN + PIM, data = goalies_cleaned)
m3 = lm(W ~ ., data = goalies_cleaned)
```

(a) Use an F-test to compares models 1 and 2. Report the following:

- The null hypothesis
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$
- The model you prefer

Solution:

```r
anova(m1, m2)
```

```
## Analysis of Variance Table
##
## Model 1: W ~ GA + SA + SV
## Model 2: W ~ GA + SA + SV + MIN + PIM
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    458 178208
## 2    456  71903  2    106305 337.09 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $H_0 : \beta_{\text{MIN}} = \beta_{\text{PIM}} = 0$
- Test statistic: $F = 337.0864493$
- P-value: $1.3395209 \times 10^{-90}$.
- Decision: **Reject** H_0 at $\alpha = 0.01$
- Model Preference: The larger model, Model 2

(b) Use an F-test to compare model 3 to your preferred model from part **(a)**. Report the following:

- The null hypothesis
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$
- The model you prefer

Solution:

```r
anova(m2, m3)
```

```
## Analysis of Variance Table
##
## Model 1: W ~ GA + SA + SV + MIN + PIM
## Model 2: W ~ GA + SA + SV + SV_PCT + GAA + SO + MIN + PIM
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    456 71903
## 2    453 70994  3    909.31 1.9341 0.1233
```

- $H_0 : \beta_{\text{MIN}} = \beta_{\text{PIM}} = 0$
- Test statistic: $F = 1.9340501$
- P-value: 0.1232638.

- Decision: **Fail to Reject** H_0 at $\alpha = 0.01$
- Model Preference: The smaller model, Model 2

(c) Use a t-test to test $H_0 : \beta_{\text{SA}} = 0$ vs $H_1 : \beta_{\text{SA}} \neq 0$ for the model you preferred in part **(b)**. Report the following:

- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.01$

Solution:

```
summary(m2)
```

```
##
## Call:
## lm(formula = W ~ GA + SA + SV + MIN + PIM, data = goalies_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.922  -3.546   1.294   2.737  63.656
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.0198636  0.7457250  -2.709 0.007011 **
## GA          -0.1359994  0.0110400 -12.319  < 2e-16 ***
## SA           0.0512308  0.0135668   3.776 0.000180 ***
## SV          -0.0581577  0.0151029  -3.851 0.000135 ***
## MIN          0.0148741  0.0006045  24.607  < 2e-16 ***
## PIM          0.0426871  0.0135557   3.149 0.001746 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.56 on 456 degrees of freedom
## Multiple R-squared:  0.9856, Adjusted R-squared:  0.9855
## F-statistic:  6262 on 5 and 456 DF,  p-value: < 2.2e-16
```

- Test statistic: $t = 3.776203$
- P-value: 1.8030573×10^{-4}
- Decision: **Reject** H_0 at $\alpha = 0.01$

Exercise 3 (Regression without `lm`)

For this exercise we will once again use the data stored in `goalies_cleaned.csv`. The goal of this exercise is to fit a model with `W` as the response and the remaining variables as predictors.

(a) Obtain the estimated regression coefficients **without** the use of `lm()` or any other built-in functions for regression. That is, you should use only matrix operations. Store the results in a vector `beta_hat_no_lm`. To ensure this is a vector, you may need to use `as.vector()`. Return this vector as well as the results of `sum(beta_hat_no_lm)`.

Solution:

```
# read in data
goalies_cleaned = read.csv("goalies_cleaned.csv")

# setup response vector and design matrix
```

```
y = goalies_cleaned$W
n = length(y)
X = cbind(rep(1, n), as.matrix(within(goalies_cleaned, rm("W"))))

# solve for estimated coefficients
beta_hat_no_lm = solve(crossprod(X, X)) %*% t(X) %*% y

# coerce results to vector
beta_hat_no_lm = as.vector(beta_hat_no_lm)

# return requested results
beta_hat_no_lm
```

```
## [1]  5.26516186 -0.11328049  0.05163855 -0.05821512 -8.04751912 -0.04960055
## [7]  0.45993589  0.01317900  0.04684216
```

```
sum(beta_hat_no_lm)
```

```
## [1] -2.431858
```

(b) Obtain the estimated regression coefficients **with** the use of `lm()`. Store the results in a vector `beta_hat_lm`. To ensure this is a vector, you may need to use `as.vector()`. Return this vector as well as the results of `sum(beta_hat_lm)`.

Solution:

```
# fit model with lm() and extract coefficients
fit = lm(W ~ ., data = goalies_cleaned)
beta_hat_lm = as.vector(coef(fit))

# return requested results
beta_hat_lm
```

```
## [1]  5.26516186 -0.11328049  0.05163855 -0.05821512 -8.04751912 -0.04960055
## [7]  0.45993589  0.01317900  0.04684216
```

```
sum(beta_hat_lm)
```

```
## [1] -2.431858
```

Notice the sum is the same, but we will complete one more step to verify equality.

(c) Use the `all.equal()` function to verify that the results are the same. You may need to remove the names of one of the vectors. The `as.vector()` function will do this as a side effect, or you can directly use `unname()`.

Solution:

```
all.equal(beta_hat_no_lm, beta_hat_lm)
```

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

Note that `all.equal()` allows for some minor differences. It only looks for "near equality." If we investigate further, we'd notice that in reality none of the estimates are the same.

```
beta_hat_no_lm == beta_hat_lm
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

But why is this? While we do know the analytic solution

$$\hat{\beta} = \left(X^\top X\right)^{-1} X^\top y,$$

it is not what `R` is using when calculating the regression coefficients using `lm()`. In reality, `R` is using a QR decomposition for speed and stability of the needed matrix operations. This creates very small numerical differences in the results.

(d) Calculate s_e without the use of `lm()`. That is, continue with your results from **(a)** and perform additional matrix operations to obtain the result. Output this result. Also, verify that this result is the same as the result obtained from `lm()`.

Solution:

```
# obtain the fitted values
y_hat = crossprod(t(X), beta_hat_no_lm)

# calculate s_e
s_e = sqrt(crossprod(y - y_hat, y - y_hat) / (length(y) - length(beta_hat_no_lm)))
s_e
```

```
##          [,1]
## [1,] 12.51873
```

```
# check that result matches lm()
all.equal(as.vector(s_e), summary(fit)$sigma)
```

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

(e) Calculate R^2 without the use of `lm()`. That is, continue with your results from **(a)** and **(d)**, and perform additional operations to obtain the result. Output this result. Also, verify that this result is the same as the result obtained from `lm()`.

Solution:

```
# some intermediate calculations
sse = sum(crossprod(y - y_hat, y - y_hat))
sst = sum(crossprod(y - mean(y), y - mean(y)))

# calculate r2
r2 = 1 - sse / sst
r2
```

```
## [1] 0.9858258
```

```
# check that result matches lm()
all.equal(as.vector(r2), summary(fit)$r.squared)
```

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

Exercise 4 (Regression for Prediction)

For this exercise use the `Boston` dataset from the `MASS` package. Use `?Boston` to learn about the dataset. The goal of this exercise is to find a model that is useful for **predicting** the response `medv`.

When evaluating a model for prediction, we often look at RMSE. However, if we both fit the model with all the data as well as evaluate RMSE using all the data, we're essentially cheating. We'd like to use RMSE as a measure of how well the model will predict on *unseen* data. If you haven't already noticed, the way we had been using RMSE resulted in RMSE decreasing as models became larger.

To correct for this, we will only use a portion of the data to fit the model, and then we will use leftover data to evaluate the model. We will call these datasets **train** (for fitting) and **test** (for evaluating). The definition of RMSE will stay the same

$$\text{RMSE}(\text{model, data}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

where

- y_i are the actual values of the response for the given data
- $\hat{y}_i$ are the predicted values using the fitted model and the predictors from the data

However, we will now evaluate it on both the **train** set and the **test** set separately. So each model you fit will have a **train** RMSE and a **test** RMSE. When calculating **test** RMSE, the predicted values will be found by predicting the response using the **test** data with the model fit using the **train** data. ***Test*** *data should never be used to fit a model.*

- Train RMSE: Model fit with train data. Evaluate on **train** data.
- Test RMSE: Model fit with train data. Evaluate on **test** data.

Set a seed of `42`, and then split the `Boston` data into two datasets, one called `train_data` and one called `test_data`. The `train_data` dataframe should contain 250 randomly chosen observations. `test_data` will contain the remaining observations. Hint: consider the following code:

```
library(MASS)
set.seed(42)
train_index = sample(1:nrow(Boston), 250)
```

Fit a total of five models using the training data.

- One must use all possible predictors.
- One must use only `tax` as a predictor.
- The remaining three you can pick to be anything you like. One of these should be the best of the five for predicting the response.

For each model report the **train** and **test** RMSE. Arrange your results in a well-formatted markdown table. Argue that one of your models is the best for predicting the response.

Solution:

```
# split the data into train and test sets
set.seed(42)
train_index = sample(1:nrow(Boston), 250) # randomly chosen observations for training
train_data  = Boston[train_index, ]
test_data   = Boston[-train_index, ]

# fit the five models
bfit_1 = lm(medv ~ tax, data = train_data)
bfit_2 = lm(medv ~ tax + ptratio + black + lstat, data = train_data)
bfit_3 = lm(medv ~ zn + nox + tax + ptratio + black + lstat, data = train_data)
bfit_4 = lm(medv ~ zn + nox + rm + dis + rad + tax + ptratio + black + lstat, data = train_data)
bfit_5 = lm(medv ~ ., data = train_data)

# function to evaluate rmse
rmse  = function(actual, predicted) {
  sqrt(mean((actual - predicted) ^ 2))
}
```

```r
# calculate all train errors
train_error = c(
  rmse(train_data$medv, predict(bfit_1, train_data)),
  rmse(train_data$medv, predict(bfit_2, train_data)),
  rmse(train_data$medv, predict(bfit_3, train_data)),
  rmse(train_data$medv, predict(bfit_4, train_data)),
  rmse(train_data$medv, predict(bfit_5, train_data))
)

# calculate all test errors
test_error = c(
  rmse(test_data$medv, predict(bfit_1, test_data)),
  rmse(test_data$medv, predict(bfit_2, test_data)),
  rmse(test_data$medv, predict(bfit_3, test_data)),
  rmse(test_data$medv, predict(bfit_4, test_data)),
  rmse(test_data$medv, predict(bfit_5, test_data))
)

model_names = c("`bfit_1`", "`bfit_2`", "`bfit_3`", "`bfit_4`", "`bfit_5`")
results = data.frame(model_names, train_error, test_error)
colnames(results) = c("Model", "Train RMSE", "Test RMSE")
knitr::kable(results)
```

Model	Train RMSE	Test RMSE
bfit_1	7.536976	8.666215
bfit_2	5.361674	6.103918
bfit_3	5.341455	6.148222
bfit_4	4.508817	5.129541
bfit_5	4.489747	5.157132

Based on these results, we believe `bfit_4` is the best model for predicting since it achieves the lowest **test** RMSE, **5.1295408**.

- `medv ~ zn + nox + rm + dis + rad + tax + ptratio + black + lstat`

First, note that the models chosen happen to be nested, but this is not necessary. However, it does illustrate that the train RMSE decreases as the size of the model increases.

Also note that the predictors for `bfit_4` were chosen in a somewhat ad-hoc manner. Consideration was given to predictors from the full model that were significant. This is not a guaranteed method, but is a decent starting point when guessing and checking.

Exercise 5 (Simulating Multiple Regression)

For this exercise we will simulate data from the following model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

Where $\epsilon_i \sim N(0, \sigma^2)$. Also, the parameters are known to be:

- $\beta_0 = 1$
- $\beta_1 = 2.5$

- $\beta_2 = 0$
- $\beta_3 = 4$
- $\beta_4 = 1$
- $\sigma^2 = 16$

We will use samples of size `n = 20`.

We will verify the distribution of $\hat{\beta}_1$ as well as investigate some hypothesis tests.

(a) We will first generate the X matrix and data frame that will be used throughout the exercise. Create the following 9 variables:

- `x0`: a vector of length `n` that contains all `1`
- `x1`: a vector of length `n` that is randomly drawn from a uniform distribution between `0` and `5`
- `x2`: a vector of length `n` that is randomly drawn from a uniform distribution between `0` and `10`
- `x3`: a vector of length `n` that is randomly drawn from a normal distribution with a mean of `0` and a standard deviation of `1`
- `x4`: a vector of length `n` that is randomly drawn from a normal distribution with a mean of `0` and a standard deviation of `2`
- `X`: a matrix that contains `x0`, `x1`, `x2`, `x3`, `x4` as its columns
- `C`: the C matrix that is defined as $(X^\top X)^{-1}$
- `y`: a vector of length `n` that contains all `0`
- `sim_data`: a data frame that stores `y` and the **four** predictor variables. `y` is currently a placeholder that we will update during the simulation

Report the diagonal of `C` as well as the 10th row of `sim_data`. For this exercise we will use the seed `1337`. Generate the above variables in the order listed after running the code below to set a seed.

```
set.seed(1337)
sample_size = 20
```

Solution:

```
x0 = rep(1, sample_size)
x1 = runif(n = sample_size, min = 0, max = 5)
x2 = runif(n = sample_size, min = 0, max = 10)
x3 = rnorm(n = sample_size, mean = 0, sd = 1)
x4 = rnorm(n = sample_size, mean = 0, sd = 2)
X = cbind(x0, x1, x2, x3, x4)
C = solve(t(X) %*% X)
y = rep(0, sample_size)
sim_data = data.frame(y, x1, x2, x3, x4)
```

```
diag(C)
```

```
##          x0          x1          x2          x3          x4
## 0.293078921 0.019753063 0.005630782 0.037903369 0.011970593
```

```
sim_data[10, ]
```

```
##    y        x1       x2         x3        x4
## 10 0 0.7302181 9.950081 0.06427374 -1.396521
```

(b) Create three vectors of length `2000` that will store results from the simulation in part **(c)**. Call them `beta_hat_1`, `beta_2_pval`, and `beta_3_pval`.

Solution:

```
num_sims    = 2000
beta_hat_1  = rep(0, num_sims)
```

```
beta_2_pval = rep(0, num_sims)
beta_3_pval = rep(0, num_sims)
```

(c) Simulate 2000 samples of size `n = 20` from the model above. Each time update the `y` value of `sim_data`. Then use `lm()` to fit a multiple regression model. Each time store:

- The value of $\hat{\beta}_1$ in `beta_hat_1`
- The p-value for the two-sided test of $\beta_2 = 0$ in `beta_2_pval`
- The p-value for the two-sided test of $\beta_3 = 0$ in `beta_3_pval`

Solution:

```
beta_0 = 1
beta_1 = 2.5
beta_2 = 0
beta_3 = 4
beta_4 = 1
sigma  = 4

library(broom)

for(i in 1:num_sims) {

  # simulate y data
  sim_data$y = with(sim_data,
                    beta_0 * x0 + beta_1 * x1 + beta_2 * x2 +
                    beta_3 * x3 + beta_4 * x4 +
                    rnorm(n = sample_size, mean = 0 , sd = sigma))

  # fit model to simulated data
  fit = lm(y ~ ., data = sim_data)

  # extract the three desired values
  beta_hat_1[i]  = coef(fit)[2]
  beta_2_pval[i] = tidy(fit)[3, ]$p.value
  beta_3_pval[i] = tidy(fit)[4, ]$p.value
}
```

(d) Based on the known values of X, what is the true distribution of $\hat{\beta}_1$?

Solution:

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 C_{11}\right)$$

$$\hat{\beta}_1 \sim N\left(\mu = 2.5, \sigma^2 = 16 \times 0.0197531 = 0.316049\right).$$

$$\hat{\beta}_1 \sim N\left(\mu = 2.5, \sigma^2 = 0.316049\right).$$

(e) Calculate the mean and variance of `beta_hat_1`. Are they close to what we would expect? Plot a histogram of `beta_hat_1`. Add a curve for the true distribution of $\hat{\beta}_1$. Does the curve seem to match the histogram?

Solution:

```r
mean(beta_hat_1)
```

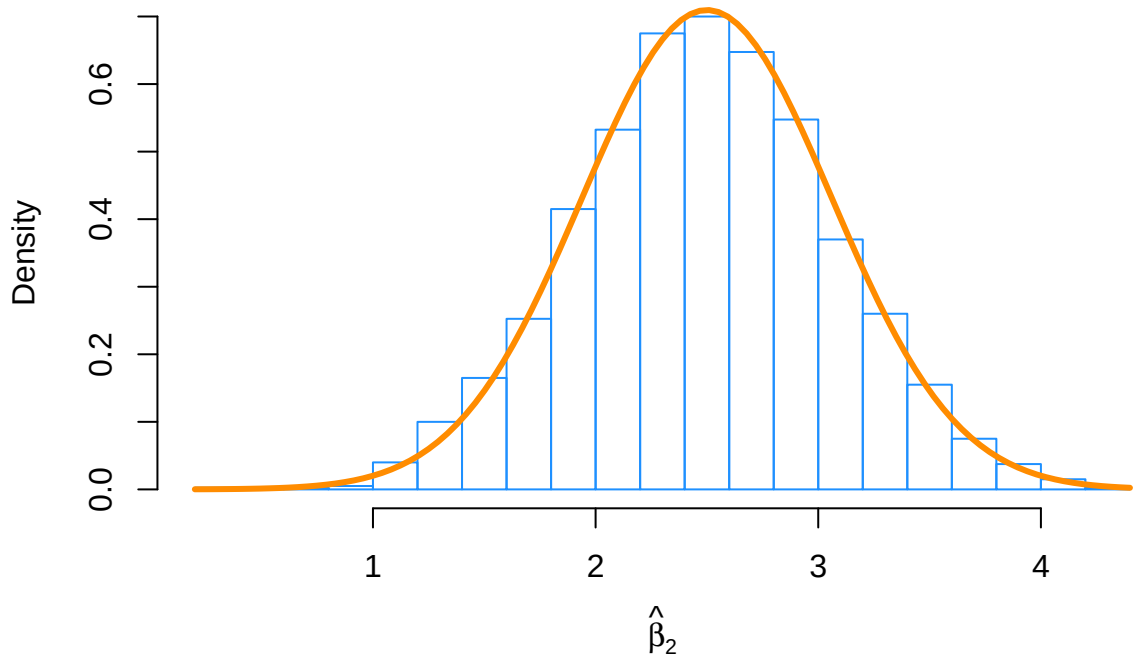
```
## [1] 2.490049
```

```r
var(beta_hat_1)
```

```
## [1] 0.3248149
```

The empirical results match what we would expect.

```r
hist(beta_hat_1, prob = TRUE, breaks = 20,
     xlab = expression(hat(beta)[2]), main = "", border = "dodgerblue")
curve(dnorm(x, mean = beta_1, sd = sqrt(sigma ^ 2 * C[1 + 1, 1 + 1])),
      col = "darkorange", add = TRUE, lwd = 3)
```



The true curve matches the histogram of simulated values well.

(f) What proportion of the p-values stored in `beta_3_pval` are less than 0.05? Is this what you would expect?

Solution:

```r
mean(beta_3_pval < 0.05)
```

```
## [1] 0.9985
```

Since $\beta_3 \neq 0$, we expect most of the p-values to be significant at $\alpha = 0.05$, so this roughly matches our expectation.

(g) What proportion of the p-values stored in `beta_2_pval` are less than 0.05? Is this what you would expect?

Solution:

```
mean(beta_2_pval < 0.05)
```

```
## [1] 0.05
```

Since $\beta_2 = 0$, we expect roughly 5% of the p-values to be significant at $\alpha = 0.05$ **by chance**, so this roughly matches our expectation. We actually see that the distribution of p-values appears roughly uniform, which is what we should expect since $\beta_2 = 0$.

```
hist(beta_2_pval, xlab = "p-value", main = "")
```