STAT 420: Homework 05

Spring 2020, Yu Wu (yuw5)

Due: Tuesday, March 3 by 11:30 PM CT

Contents

Assignment **1**
Exercise 1 (Using `lm`) 1
Exercise 2 (More `lm`) 4
Exercise 3 (Comparing Models) 6
Exercise 4 (Regression without `lm`) 9

Assignment

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.

```
nutrition <- read.csv("nutrition.csv")

fit <- lm(Calories ~ Carbs + Fat + Protein, data = nutrition)
summary(fit)
```

```
##
## 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
```

- The null hypotheses is that all of the predictors are insignificant. The alternative hypothesis is that at least one of the predictors is significant.
- The value of the test statistic is 1.524e05.
- The p-value of the test is 2.2e-16.
- A statistical decision at $\alpha = 0.01$ is reject the null hypothesis since the p-value is smaller than alpha.
- A conclusion in the context of the problem is that the amount of calories of a food is related with at least one of the following predictors: Carbs, Fat and Protein.

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

```
summary(fit)$coef[,1]
```

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

$$Calories = 3.768 + 3.774 * Carbs + 8.804 * Fat + 3.967 * Protein$$

When there is no carbs, fat or protein, the amount of calories is 3.768; when carbs increases by one, the amount of calories increases by 3.774; when fat increases by one, the amount of calories increases by 8.804; when protein increases by one, the amount of calories increases by 3.967;

(c) Use your model to predict the amount 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.

```
3.768 + 3.774*47 + 8.804*28 + 3.967*25
```

```
## [1] 526.833
```

$$Calories_{BIGMAC} = 3.768 + 3.774 * 47 + 8.804 * 28 + 3.967 * 25 = 526.833$$

Big Mac has 526.833g calories.

(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.

```
s_y <- sd(nutrition$Calories)
s_e <- summary(fit)$sigma

s_y
```

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

```
s_e
```

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

The standard deviation tells us how much the calories spread out within the dataset. The standard error tells us how much our residuals or the fitted values spread out within the fitted model.

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

```
summary(fit)$r.squared
```

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

The R^2 of the model is 0.989. It means 98.9% of the data is explained by the fitted model.

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

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

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

We are 90% confident that the value of β_2 is between 8.779 and 8.829.

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

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

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

We are 95% confident that the value of β_0 is between 2.803 and 4.733.

(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.

```
x <- data.frame(Carbs = 30, Fat = 11, Protein = 2)
predict(fit, newdata = x, interval = 'confidence', level = 0.99)
```

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

We are 99% sure that the amount of calories of french fries with the given data is between 220.892g and 222.620g.

(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.

```
x <- data.frame(Carbs = 11, Fat = 1.5, Protein = 1)
predict(fit, newdata = x, interval = 'confidence', level = 0.90)
```

```
##        fit      lwr      upr
## 1 62.45115 61.77276 63.12954
```

We are 90% sure that the amount of calories of french fries with the given data is between 61.77276g and 63.12954g.

Exercise 2 (More `lm`)

For this exercise we will again use the nutrition data.

(a) Fit a model with Calories as the response and `Carbs`, `Sodium`, `Fat`, and `Protein` as predictors. 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

```
fit <- lm(Calories ~ Carbs + Sodium + Fat + Protein, data = nutrition)
summary(fit)
```

```
##
## Call:
## lm(formula = Calories ~ Carbs + Sodium + Fat + Protein, data = nutrition)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -381.00   -3.46   -0.32    5.30  259.31
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.6930721  0.4954121   7.455 1.05e-13 ***
## Carbs       3.7730437  0.0097064 388.717  < 2e-16 ***
## Sodium      0.0003305  0.0002425   1.363    0.173
## Fat         8.8039259  0.0153042 575.262  < 2e-16 ***
## Protein     3.9649053  0.0263390 150.533  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.89 on 5133 degrees of freedom
## Multiple R-squared:  0.9889, Adjusted R-squared:  0.9889
## F-statistic: 1.144e+05 on 4 and 5133 DF,  p-value: < 2.2e-16
```

- The null hypotheses is all of the predictors are
- 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

(b) For each of the predictors in part **(a)**, perform a t-test for the significance of its regression coefficient. Report the following for each:

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

```
summary(fit)$coef
```

```
##                 Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) 3.6930721445 0.4954121127   7.454546 1.052621e-13
## Carbs       3.7730437360 0.0097063959 388.717271 0.000000e+00
## Sodium      0.0003304667 0.0002425302   1.362579 1.730749e-01
## Fat         8.8039259458 0.0153042101 575.261702 0.000000e+00
## Protein     3.9649052931 0.0263390430 150.533385 0.000000e+00
```

Carbs:

- $H_0 : \beta_{Carbs} = 0$ VS $H_a : \beta_{Carbs} \neq 0$
- $t = 388.717271$
- $p - value = 0$
- A statistical decision at $\alpha = 0.01$ is reject the null.

Sodium:

- $H_0 : \beta_{Sodium} = 0$ VS $H_a : \beta_{Sodium} \neq 0$
- $t = 1.362579$
- $p - value = 1.730749e - 01$
- A statistical decision at $\alpha = 0.01$ is fail to reject the null.

Fat:

- $H_0 : \beta_{Fat} = 0$ VS $H_a : \beta_{Fat} \neq 0$
- $t = 575.261702$
- $p - value = 0$
- A statistical decision at $\alpha = 0.01$ is reject the null.

Protein:

- $H_0 : \beta_{Protein} = 0$ VS $H_a : \beta_{Protein} \neq 0$
- $t = 150.533385$
- $p - value = 0$
- A statistical decision at $\alpha = 0.01$ is reject the null.

(c) Based on your results in part **(b)**, do you still prefer the model in part **(a)**, or is there instead a model with three predictors that you prefer? Briefly explain.

No, because the Sodium predictor is insignificant, and we can drop it from the model. The reduced model would be the one in part 1(a).

Exercise 3 (Comparing Models)

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

(a) Fit a multiple linear regression model with Wins as the response and all other variables as the predictors.

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 $\alpha = 0.10$
- 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.

```
goalies_cleaned <- read.csv("goalies_cleaned.csv")

fit <- lm(W ~ ., data = goalies_cleaned)
summary(fit)
```

```
##
## Call:
## lm(formula = W ~ ., data = goalies_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.204  -3.126   0.935   2.835  64.078
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.2651619 16.8181423   0.313 0.754376
## GA          -0.1132805  0.0148085  -7.650 1.22e-13 ***
## SA           0.0516385  0.0135565   3.809 0.000159 ***
## SV          -0.0582151  0.0150905  -3.858 0.000131 ***
## SV_PCT      -8.0475191 17.6600154  -0.456 0.648830
## GAA         -0.0496006  0.4821957  -0.103 0.918116
## SO           0.4599359  0.1989567   2.312 0.021240 *
## MIN          0.0131790  0.0009504  13.867  < 2e-16 ***
## PIM          0.0468422  0.0136373   3.435 0.000647 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.52 on 453 degrees of freedom
## Multiple R-squared:  0.9858, Adjusted R-squared:  0.9856
## F-statistic:  3938 on 8 and 453 DF,  p-value: < 2.2e-16
```

- $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$ VS $H_a : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 \neq 0$
- $F = 3938$
- $p - value = 2.2e - 16$
- A statistical decision at $\alpha = 0.10$ is reject the null.
- Based on the F-Test, we can conclude that the number of wins is related with at least one of those predictors.

(b) Calculate the RMSE of this full model. Report the residual standard error of this full model. What is the relationship of these two values?

Recall, we have defined RMSE as,

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

```
rmse <- sqrt(mean(resid(fit)^2))
rmse
```

```
## [1] 12.3962
```

```
summary(fit)$sigma
```

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

The two values are both estimators of standard deviation of errors.

(c) Fit a model with Wins as the response and with Goals Against, Goals Against Average, Saves, and Save Percentage as the predictors. Calculate the RMSE of this model.

```
fit2 <- lm(W ~ GA + GAA + SV + SV_PCT, data = goalies_cleaned)
summary(fit2)
```

```
##
## Call:
## lm(formula = W ~ GA + GAA + SV + SV_PCT, data = goalies_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -88.101  -8.637   1.912   5.469 183.405
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.066e+01  3.386e+01   0.610    0.542
## GA           9.645e-02  4.106e-03  23.487   <2e-16 ***
## GAA         -1.876e+00  9.673e-01  -1.940    0.053 .
## SV           8.389e-03  4.933e-04  17.005   <2e-16 ***
## SV_PCT      -2.166e+01  3.555e+01  -0.609    0.543
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 25.26 on 457 degrees of freedom
## Multiple R-squared:  0.9418, Adjusted R-squared:  0.9413
## F-statistic:  1848 on 4 and 457 DF,  p-value: < 2.2e-16
```

```
rmse2 <- sqrt(mean(resid(fit2)^2))
rmse2
```

```
## [1] 25.12237
```

(d) Fit a model with Wins as the response and with Goals Against Average and Save Percentage as the predictors. Calculate the RMSE of this model.

```
fit3 <- lm(W ~ GAA +SV_PCT, data = goalies_cleaned)
summary(fit3)
```

```
##
## Call:
## lm(formula = W ~ GAA + SV_PCT, data = goalies_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -109.88  -65.45  -44.63   38.02  608.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -98.804    136.442  -0.724    0.469
## GAA           -4.242      3.947  -1.075    0.283
## SV_PCT       208.682    142.817   1.461    0.145
##
## Residual standard error: 103.2 on 459 degrees of freedom
## Multiple R-squared:  0.02464,    Adjusted R-squared:  0.02039
## F-statistic: 5.797 on 2 and 459 DF,  p-value: 0.003264
```

```
rmse3 <- sqrt(mean(resid(fit3)^2))
rmse3
```

```
## [1] 102.8307
```

(e) Based on the previous three models, which model is most helpful for predicting wins? Briefly explain.

The full model is most helpful amoung all the others, because it has the lowest RMSE.

(f) Conduct an ANOVA F-test comparing the models in parts **(c)** and **(d)**. Report the following:

- The null and alternative hypotheses
- The value of the test statistic
- The p-value of the test
- A statistical decision at $\alpha = 0.10$
- 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.

```
anova(fit3, fit2)
```

```
## Analysis of Variance Table
##
## Model 1: W ~ GAA + SV_PCT
## Model 2: W ~ GA + GAA + SV + SV_PCT
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
```

```
## 1    459 4885261
## 2    457  291584  2   4593677 3599.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $H_0 : \beta_{GA} = \beta_{SV} = 0$ VS $H_a : \beta_{GA} = \beta_{SV} \neq 0$
- $F=3599.8 $
- $p - value = 2.2e - 16$
- A statistical decision at $\alpha = 0.10$ is reject the null.
- Based on the F-Test, we can conclude that the model that contains GA and SV is better.

Exercise 4 (Regression without `lm`)

For this exercise use the `prostate` dataset from the `faraway` package. Use `?prosate` to learn about the dataset. The goal of this exercise is to fit a model with `lpsa` 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)`.

```
library("faraway")
data("prostate")

n = nrow(prostate)
x = as.matrix(cbind(rep(1, n), prostate[,1-ncol(prostate)-1]))
y = prostate$lpsa
beta_hat_no_lm = as.vector(solve(t(x) %*% x) %*% t(x) %*% y)
beta_hat_no_lm
```

```
## [1]  0.669336698  0.587021826  0.454467424 -0.019637176  0.107054031
## [6]  0.766157326 -0.105474263  0.045141598  0.004525231
```

```
sum(beta_hat_no_lm)
```

```
## [1] 2.508593
```

(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)`.

```
fit <- lm(lpsa ~ ., data = prostate)
beta_hat_lm <- as.vector(summary(fit)$coef[, 1])
beta_hat_lm
```

```
## [1]  0.669336698  0.587021826  0.454467424 -0.019637176  0.107054031
## [6]  0.766157326 -0.105474263  0.045141598  0.004525231
```

```
sum(beta_hat_lm)
```

```
## [1] 2.508593
```

(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()`.

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

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

(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()`.

```
p <- length(coef(fit))
y_hat <- x %*% solve(t(x) %*% x) %*% t(x) %*% y
e <- y - y_hat
s_e <- sqrt(t(e) %*% e / (n - p))

s_e
```

```
##           [,1]
## [1,] 0.7084155
```

```
all.equal(summary(fit)$sigma,as.vector(s_e))
```

```
## [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()`.

```
SST <- sum((y - mean(y)) ^ 2)
SSF <- sum((y_hat - mean(y)) ^ 2)
R2 <- SSF / SST

R2
```

```
## [1] 0.6547541
```

```
all.equal(R2,summary(fit)$r.squared)
```

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