Assignment

John Karuitha

January 11, 2024

Contents

1 Question One

Here are average systolic blood pressure data for children/young adults aged 1 through 19 (we used these exact data in homework 8). Run the following lines of code:

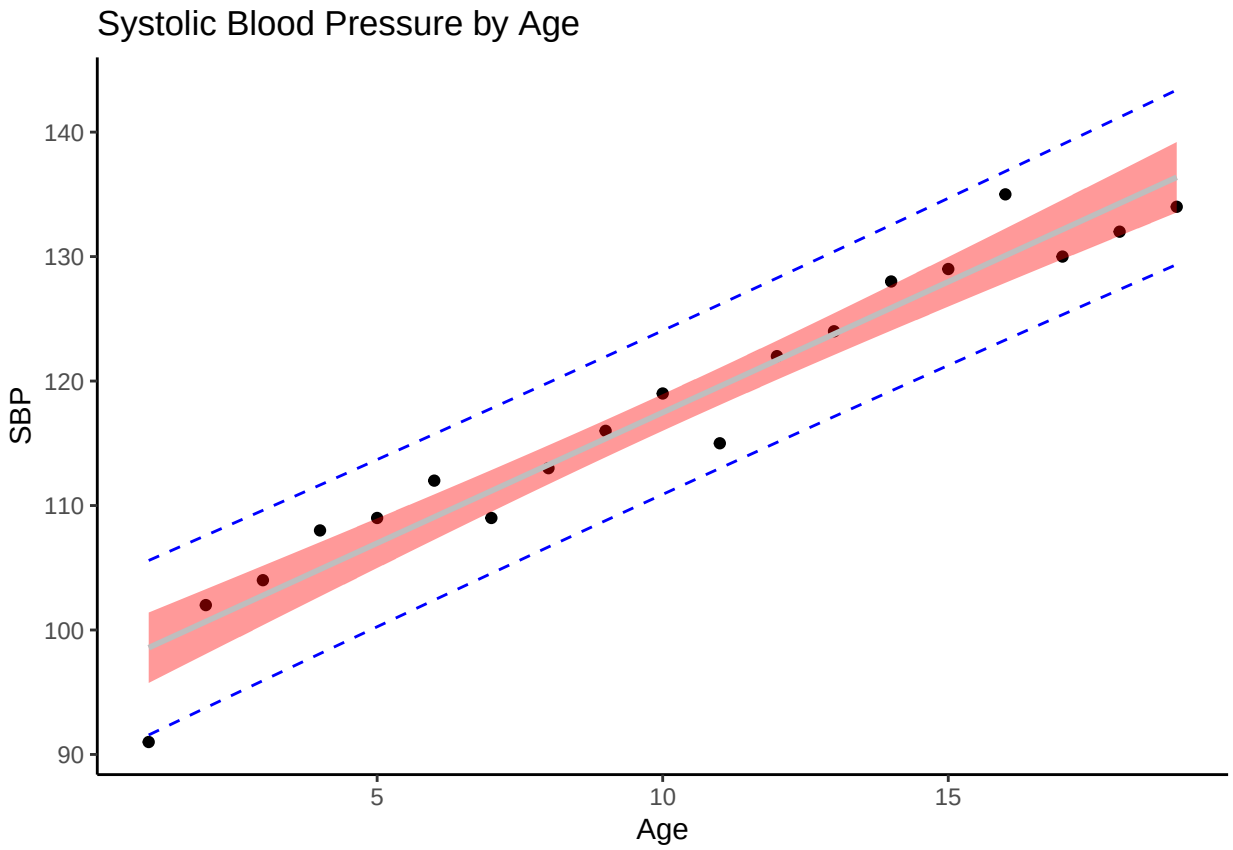
```
age <- c(1:19)
SBP <- c(91,102,104,108,109,112,109,113,116,119,115,122,124,128,129,135,130,132,134)
bpstudy <- data.frame(age,SBP)
```

1.1 Reproduce the scatter plot from homework 8, question 9

This time including the prediction intervals (make these blue), and the confidence intervals (make these red). As a reminder, the homework 8 question asked you to create a scatterplot showing both the individual data points and the best regression line fit through them, and to add labels to both axes and show the data points using a different color than the line.

```
## Create a model
my_reg_model <- lm(SBP ~ age, data = bpstudy)
##create a prediction
prediction <- predict(my_reg_model, interval = "prediction")
prediction <- as.data.frame(prediction)
bpstudy <- bpstudy %>%
    bind_cols(prediction)

##Plot the data
bpstudy %>%
    ggplot(mapping = aes(x = age)) +
    geom_point(aes(y = SBP)) +
    geom_smooth(aes(y = SBP), fill = "red", method = "lm", col = "gray") +
    geom_line(aes(y = lwr), linetype = "dashed", col = "blue") +
    geom_line(aes(y = upr), linetype = "dashed", col = "blue") +
    labs(x = "Age", y = "SBP", title = "Systolic Blood Pressure by Age")
```



1.2 Explain the difference between the `confidence intervals` and the `prediction intervals` in words.

The confidence interval shows the possible range of values associated with a given parameter such as the mean or median. In other words, what range of values in the sample data is likely to contain a parameter of interest such as the mean or median?

On the other hand, a prediction interval is the interval or range within which a future observation or set of observation is likely to fall. Hence, the prediction interval informs us about the preciseness of our forecasts.

1.3 Do there appear to be any outliers in these data? Use the definition of outlier as exceeding more than 1.5 times the interquartile range in either direction. How do these point(s) appear on the scatterplot?

There appears to be no outliers in the plot.

```
my_iqr <- IQR(bpstudy$SBP) * 1.5

q3 <- quantile(bpstudy$SBP, probs = 0.75)

q1 <- quantile(bpstudy$SBP, probs = 0.25)

upper_bound = q3 + my_iqr
lower_bound = q1 - my_iqr

bpstudy$SBP > upper_bound
```

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

```
## [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
bpstudy$SBP < lower_bound
```

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

1.4 Rerun the model after removing any outliers. (If there were no outliers, remove the single point with the highest residual). How does the regression model change? Consider coefficients and R-squared.

Because there are no outliers, I remove the point with the highest residual, which happens to be the first point in the dataset.

```
residuals(my_reg_model) %>% abs() %>% max()
```

```
## [1] 7.573684
```

```
new_data <- bpstudy[-1, ]


new_reg <- lm(SBP ~ age, data = new_data)

stargazer::stargazer(my_reg_model, new_reg, type = "text")
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                         SBP
##                              (1)                    (2)
## ---------------------------------------------------------------
## age                        2.100***               1.951***
##                            (0.127)                (0.105)
##
## Constant                   96.474***              98.454***
##                            (1.451)                (1.232)
##
## ---------------------------------------------------------------
## Observations                 19                     18
## R2                          0.941                  0.956
## Adjusted R2                 0.938                  0.953
## Residual Std. Error     3.039 (df = 17)        2.316 (df = 16)
## F Statistic         272.120*** (df = 1; 17) 344.064*** (df = 1; 16)
## ===============================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

The R^2 and adjusted R^2 do improve significantly after removing the most extreme value in the data set. The coefficient for age rises while the constant in the regression equation goes down.

2 Question Two

Attached to this assignment is `health-status.csv` a clean copy of the 2020 NHIS file used throughout the semester. This version has the variables already renamed and recorded to NA as required. The aim is to explore regression models that predict self-reported health status as a function of age, sex, income, education, and alcohol consumption. Here are brief summaries of the values.

You can consult the data dictionary for more detail where necessary.

- age - originally AGEP_A - age in years and 85+
- sex - originally SEX_A - 1=male, 2=female
- education - originally EDUC_A - 00 - never attended school up to 11 - doctoral degree
- alcohol - average number of days per week in the past year in which alcoholic beverages were consumed from 0 to 7
- income - originally INCGRP_A - family income with categories of 1 (lowest) to 5 (highest)
- health - general health status 1-excellent, 2-very good, 3-good, 4-fair, 5-poor

```
health <- read_csv("health-status.csv") %>%
    select(-`...1`, -X)
```

2.1 Fit a regression model relating health status to age, sex, income, education, and alcohol consumption.

```
health_reg <- lm(health ~ ., data = health)
summary(health_reg)
```

```
##
## Call:
## lm(formula = health ~ ., data = health)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4350 -0.7539 -0.0265  0.6520  3.4127
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.6212024  0.0294421  89.029  < 2e-16 ***
## age          0.0124396  0.0003078  40.414  < 2e-16 ***
## sex         -0.0557810  0.0111688  -4.994 5.93e-07 ***
## income      -0.1320181  0.0039370 -33.533  < 2e-16 ***
## education   -0.0632032  0.0024944 -25.338  < 2e-16 ***
## alcohol     -0.0343176  0.0029872 -11.488  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9673 on 31085 degrees of freedom
##   (477 observations deleted due to missingness)
## Multiple R-squared:  0.158,  Adjusted R-squared:  0.1578
## F-statistic:  1166 on 5 and 31085 DF,  p-value: < 2.2e-16
```

2.2 Interpret each of the beta coefficients in terms of the direction and magnitude of the predictions. For example, if body mass index (BMI) was in the model with a coefficient of 0.05, we would say that each unit increase in BMI reduced the health score by an average of 0.05, meaning that BMI is inversely correlated with general health.

- β_0: β_0 is the baseline health level given that a person cannot have 0 (perfect health) or 6 (no health) unless the person is superhuman or dead.
- β_1: β_1 for `age` is 0.0124396. This coefficient is positive meaning that as the respondents in the sample increased in age, the poorer their health status. Specifically, for each unit rise in age, health tends to

deteriorate by 0.0124396 units (with health on a scale of 1 to 5) all other variables remaining the same.

- β_2: β_2 of -0.0557810 corresponds to sex, with 1 being male and 2 for female. The coefficient shows that holding all other variables constant, women have less health issues compared to men. For instance if you got a man and a woman with the same characteristics except for their different sex, then the man is likely to be 0.0557810 units less healthy than the woman on a scale of 1 to 5 for health.
- β_3: Income has a coefficient of -0.1320181 meaning that more income corresponds to better health. A unit increase in income raise the health metric by 0.1320181 on a scale of 1 (excellent) to 5 (poor), all else remaining the same.
- β_4: Education has a coefficient of -0.0632032 meaning that more education corresponds to better health, ceteris paribus. Again a unit rise in education raises health by 0.0632032 on a scale of 1 (excellent health) to 5 (poor health).
- β_5: Contrary to research, Alcohol, with a coeffcient of -0.0343176 implies that the more alcohol one consumes, the better the health. An extra day of drinking corresponds to an improvement of health by 0.0343176 on the scale of 1 to 5.

2.3 Does the result for alcohol consumption agree with or defy your expectations? Propose an alternative way of measuring alcohol consumption that might better capture the known health effects of heavy alcohol use.

The result for alcohol does not make sense. While minimum amounts of alcohol could improve health, daily drinking certainly does not. The number of days a person drinks may not always capture the quantity of alcohol consumed.

We could improve this model by measuring alcohol consumption in terms of quantity of alcohol consumes per week or per day.

2.4 According to this model, what is the predicted health score for a 45-year old male bachelor's degree holder with a family income of $88,000 who consumes alcohol daily?

```
my_vars <- data.frame(age = 45, sex = 1, income = 4, education = 8, alcohol = 7)

predict(health_reg, newdata = my_vars)
```

```
##        1
## 1.851283
```

2.5 According to this model, what type of person (in terms of age, sex, income, education, and alcohol consumption) would be expected to have the best health score? The worst?

According to this model, a person with the best health shoud be:

- Age: Young (18 is the youngest in this case).
- Sex: Female.
- Income: Highest Income Band (level 5).
- Education: Highest Education Level (11: Doctoral degree).
- Alcohol: Drinks everyday (7).

2.6 Someone reviews these model results and says "You are telling me that just because I am 75 years old I likely have poor health. But I rate my health as excellent. This model is worthless". How would you respond to this criticism?

The response should be as follows.

If we had a sample of younger people who had similar characteristics with the 75 year old man (except for age), then these younger people would, on average, have better health scores than the 75 year old, all else remaining the same.