Homework 6

Kevin Linares and Jamila Sani

2024-10-24

```
require(ISLR2)
library(faraway)
library(knitr)
library(tidyverse)

options(scipen=999)
```

1.Faraway Chapter 4. Exercise 2. Using the teengamb dataset from the faraway package Fit a linear regression model with gamble as the outcome and the other variables

- We model gamble, expenditure on gambling in pounds per year, as a function on sex (male, female), status score based on parents' occupation), income (in pounds per week), and verbal scores (in words out of 12 correctly defined).

```
data("teengamb")
?teengamb
glimpse(teengamb)
```

```
Rows: 47
Columns: 5
$ sex    <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, ~
$ status <int> 51, 28, 37, 28, 65, 61, 28, 27, 43, 18, 18, 43, 30, 28, 38, 38,~
$ income <dbl> 2.00, 2.50, 2.00, 7.00, 2.00, 3.47, 5.50, 6.42, 2.00, 6.00, 3.0~
$ verbal <int> 8, 8, 6, 4, 8, 6, 7, 5, 6, 7, 6, 6, 4, 6, 6, 8, 8, 5, 8, 9, 8, ~
$ gamble <dbl> 0.00, 0.00, 0.00, 7.30, 19.60, 0.10, 1.45, 6.60, 1.70, 0.10, 0.~
```

```r
summary(
  mod <- lm(gamble ~ sex + status + income + verbal, teengamb))
```

```
Call:
lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-51.082 -11.320  -1.451   9.452  94.252

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  22.55565   17.19680   1.312    0.1968
sex         -22.11833    8.21111  -2.694    0.0101 *
status        0.05223    0.28111   0.186    0.8535
income        4.96198    1.02539   4.839 0.0000179 ***
verbal       -2.95949    2.17215  -1.362    0.1803
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.69 on 42 degrees of freedom
Multiple R-squared:  0.5267,    Adjusted R-squared:  0.4816
F-statistic: 11.69 on 4 and 42 DF,  p-value: 0.000001815
```

- A. Predict the amount that a male with average values (given these data) for status, income, and verbal would spend on gambling along with an appropriate 95% interval.
  - We would expect a male with average SES status, income, and verbal scores to spend on average 28.24 pounds per week on gambling, and we are 95% confident taht the next observation with these predictor values will fall between -18.52 to 75.00 pounds per week.

```r
male_mean_pred <- teengamb |>
  summarise(
    status = mean(status),
    income = mean(income),
    verbal = mean(verbal)) |>
  mutate(sex = 0) |> relocate(sex)

male_mean_pred |> kable()
```

sex	status	income	verbal
0	45.23404	4.641915	6.659574

```
predict(mod,
        newdata = male_mean_pred, interval = "prediction") |>
  kable()
```

fit	lwr	upr
28.24252	-18.51536	75.00039

- B. Repeat the prediction for a male with maximal values (for these data) of status, income, and verbal. Which interval is wider and why is this result expected?
  - For a male with max values of status, income, and verbal scores we would expect to spend on average 71.3 pounds per week on gambling, with a 95% prediction interval between 17.1 and 125.6.
  - Among a male with max scores, the difference between the upper and lower 95% prediction interval is 16% wider (108.5) versus a male with average scores (93.5), suggesting that the the wider interval for males with max scores corresponds with more uncertainty in the model's prediction of these individuals due to the variability (random error) in the data. It is likely that our data does not contain many observations near the max scores, and thus our predictions become more uncertain as we move away from the mean of scores.

```
male_max_pred <- teengamb |>
  summarise(
    status = max(status),
    income = max(income),
    verbal = max(verbal)) |>
  mutate(sex = 0) |> relocate(sex)

male_max_pred |> kable()
```

sex	status	income	verbal
0	75	15	10

```r
predict(mod,
        newdata = male_max_pred, interval = "prediction") |>
  kable()
```

fit	lwr	upr
71.30794	17.06588	125.55

- C. Fit a model with sqrt (gamble) as the outcome but with the same predictors. Now predict the response and estimate the appropriate 95% interval for the individual in #A. Take care to give your answer on the original scale of the response (i.e., gamble not sqrt (gamble).
  - We would expect a male with average SES status, income, and verbal scores to spend on average 16.40 pounds per week on gambling, and we are 95% confident taht the next observation with these predictor values will fall between 0.06 and 69.62 pounds per week.

```r
summary(
  mod2 <- lm(sqrt(gamble) ~ sex + status + income + verbal, teengamb))
```

```
Call:
lm(formula = sqrt(gamble) ~ sex + status + income + verbal, data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6606 -1.0961 -0.2564  0.9786  5.4178

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept)  2.97707    1.57947   1.885    0.06638 .
sex         -2.04450    0.75416  -2.711    0.00968 **
status       0.03688    0.02582   1.428    0.16057
income       0.47938    0.09418   5.090 0.00000794 ***
verbal      -0.42360    0.19950  -2.123    0.03967 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.084 on 42 degrees of freedom
Multiple R-squared:  0.5646,    Adjusted R-squared:  0.5231
F-statistic: 13.61 on 4 and 42 DF,  p-value: 0.0000003362
```

```
predict(mod2,
        newdata = male_mean_pred, interval = "prediction")^2 |>
  kable()
```

fit	lwr	upr
16.39864	0.0600422	69.6237

- D. Repeat the prediction for the model in #C for a female with status=20, income=1, verbal=10. Give your answer on the original scale of the response and comment on the credibility of the result.
  - We would expect a female with status of 20, income of 1, and verbal score of 10 to spend on average 4.35 pounds per week on gambling, and we are 95% confident taht the next observation with these predictor values will fall between 47.73 and 7.49 pounds per week. The prediction falls outside of the prediction interval after transforming back from the square root scale. Additionally, the lower bound is much higher than the upper band and we conclude that predictions for some people are not trustworthy and perhaps modeling the square root of the outcome is not feasible.

```
female_pred <- teengamb |>
  summarise(
    status = 20,
    income = 1,
    verbal = 10) |>
  mutate(sex = 1) |> relocate(sex)

predict(mod2,
        newdata = female_pred, interval = "prediction")^2 |>
  kable()
```

fit	lwr	upr
4.353398	47.73238	7.485167

- E. Now, go back to #A. Consider how much males with average values (given these data) for status, income, and verbal would spend on gambling along with an appropriate 95% interval. How does this interval compare to the interval in #A? Explain whether this is expected and why.

- We construct a confidence interval for male with average scores, to suggest that if we were to take repeated samples we would expect the 95% CIs of the samples to contain the true mean 95% of the time. Our confidence interval is different from the prediction interval reported in question A., given that a prediction interval is a range of values that is likely to contain a single new observation based on a specified subset of predictor values. Our confidence interval is much more narrower than the prediction value. The difference is that with confidence intervals we assess variability around the mean in an estimated quantity on the predictor, while the prediction interval looks at a single new observation, not the average, so we have to account for the variability of the predictor in addition to the variability of our estimate of the mean.

```
predict(mod,
        newdata = male_mean_pred, interval = "confidence") |>
  kable()
```

fit	lwr	upr
28.24252	18.78277	37.70227

2. Assess the following DAG. You want to get an unbiased estimate for the association between HIV and Stroke. What covariate(s) would you want to include in your model and why?

- In our model to get an unbiased estimate between HIV and Stroke we would not include any covariate as it may introduce post-treatment bias, concluding that HIV does not have an effect on stroke. In the DAG we have a pipe covariate from age -> smoking -> HIV -> stroke. Therefore, we would be stratifying HIV by smoking, and smoking by age thus blocking the path between HIV and stroke and thus statistically removed from the results.