HW 6

Evan Cranmer and Rohin Mishra

2024-10-23

Question 1a

```
library(faraway)
data(teengamb)

#familiarizing ourselves with the dataset
?teengamb

gamblemodel <- lm(gamble ~ ., data=teengamb)
summary(gamblemodel)
```

```
##
## Call:
## lm(formula = gamble ~ ., 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 1.79e-05 ***
## 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: 1.815e-06
```

```
#averages for males
averages <- teengamb %>%
  filter(sex == 0) %>%
  summarise(
    avg_status = mean(status),
    avg_income = mean(income),
    avg_verbal = mean(verbal)
```

```
  )

#putting it in a data frame
averages_male <- data.frame(
  status = averages$avg_status,
  income = averages$avg_income,
  verbal = averages$avg_verbal,
  sex = 0
)

predictions_1a <- predict(
  gamblemodel,
  newdata = averages_male,
  type = 'response',
  se = TRUE
)

#calculating our 95% confidence interval
quant <- qt(0.025, nrow(teengamb) - ncol(averages_male), lower.tail = FALSE)
sigma2 <- sum(residuals(gamblemodel)^2) / (nrow(teengamb) - 4)

#calculating standard error
se.fit.pred <- sqrt(sigma2 * (1 + (predictions_1a$se.fit^2 / sigma2)))

df_preds_pred <- data.frame(averages_male, predictions_1a) %>%
  mutate(
    lower = fit - (quant * se.fit.pred),
    upper = fit + (quant * se.fit.pred)
  ) %>%
  summarise(
    mn_beta = mean(fit),
    lcl = mean(lower),
    ucl = mean(upper)
  )

print(df_preds_pred)
```

```
##   mn_beta      lcl     ucl
## 1  29.775 -16.2686 75.8186
```

Our model predicts that the average expenditure on gambling in pounds per year is 29.77. Our 95% interval is (-16.2686, 75.8186).

Question 1b repeating for max values

```
#predicting max value for males

max_values <- teengamb %>%
  filter(sex == 0) %>%
  summarise(
    max_status = max(status),
```

```
    max_income = max(income),
    max_verbal = max(verbal)
  )

datnew_max <- data.frame(
  status = max_values$max_status,
  income = max_values$max_income,
  verbal = max_values$max_verbal,
  sex = 0
)

#getting our prediction intervals
predictions_max <- predict(
  gamblemodel,
  newdata = datnew_max,
  type = 'response',
  se = TRUE
)

#getting our confidence intervals
se.fit.pred_max <- sqrt(sigma2 * (1 + (predictions_max$se.fit^2 / sigma2)))

df_preds_pred_max <- data.frame(datnew_max, predictions_max) %>%
  mutate(
    lower = fit - (quant * se.fit.pred_max),
    upper = fit + (quant * se.fit.pred_max)
  ) %>%
  summarise(
    mn_beta = mean(fit),
    lcl = mean(lower),
    ucl = mean(upper)
  )

#viewing our results
print(df_preds_pred_max)
```

```
##    mn_beta      lcl      ucl
## 1 71.30794 17.55429 125.0616
```

Our estimate for a male with maximal values is spending 71.3077 pounds per year on gambling. Our 95% interval is (17.55429, 125.0616). This prediction interval is wider because linear regression assumes homoscedasticity, but predictions at the maximum involve greater variability, which would make the interval wider than our average value interval interval.

Question 1c

```
#fitting our model
sqrt_gamble <- lm(sqrt(gamble) ~ status + income + verbal + sex, data = teengamb)
sigma2 <- sum(residuals(sqrt_gamble)^2) / (nrow(teengamb) - 4)

predictions_sqrt <- predict(
```

```
  sqrt_gamble,
  newdata = averages_male,
  se = TRUE
)

#getting fit and se
fit_sqrt <- predictions_sqrt$fit
se.fit <- predictions_sqrt$se.fit

#95% CIs
quant <- qt(0.025, nrow(teengamb) - 4, lower.tail = FALSE)

#calculating lower and upper bounds on the square root scale
lower_sqrt <- fit_sqrt - (quant * se.fit)
upper_sqrt <- fit_sqrt + (quant * se.fit)

#CONVERTING BACK TO ORIGINAL
mn_beta_original <- fit_sqrt^2
lcl_original <- lower_sqrt^2
ucl_original <- upper_sqrt^2

df_preds_pred_original <- data.frame(
  mn_beta = mn_beta_original,
  lcl = lcl_original,
  ucl = ucl_original
)

print(df_preds_pred_original)
```

```
##    mn_beta      lcl      ucl
## 1 19.27805 12.93419 26.88363
```

Our estimate for an average male in this model is spending 19.27805 pounds per year on gambling. Our 95% CI is (12.93419, 26.88363).

Question 1d

```
#storing our new values for our individual
female_values <- data.frame(
  status = 20,
  income = 1,
  verbal = 10,
  sex = 1
)

predictions_female_sqrt <- predict(
  sqrt_gamble,
  newdata = female_values,
  type = 'response',
  se = TRUE
)
```

```r
fit_sqrt_female <- predictions_female_sqrt$fit
se.fit_female <- predictions_female_sqrt$se.fit
quant <- qt(0.025, nrow(teengamb) - 4, lower.tail = FALSE)

#calculating our CIs
lower_sqrt_female <- fit_sqrt_female - (quant * se.fit_female)
upper_sqrt_female <- fit_sqrt_female + (quant * se.fit_female)

#CONVERTING TO OG SCALE
mn_beta_original_female <- fit_sqrt_female^2
lcl_original_female <- lower_sqrt_female^2
ucl_original_female <- upper_sqrt_female^2

female_original <- data.frame(
  mn_beta = mn_beta_original_female,
  lcl = lcl_original_female,
  ucl = ucl_original_female
)

print(female_original)
```

```
##    mn_beta      lcl       ucl
## 1 4.353398 19.75192 0.0736327
```

The prediction for how much a female with these values would spend on gambling per year is 4.353398 pounds. The stated confidence interval has a *lower* bound of 19.76636 and *upper* bound of 0.07451699, which does not make sense. Before our transformation back to the original scale, the lower bound had a negative value further away from zero than our upper bound, but when we transform this, it ends up being a greater positive number.

Question 1e: Going off of observed values.

```r
teengamb_males <- teengamb %>% filter(sex == 0)

#getting our averages
avg_status <- mean(teengamb_males$status)
avg_income <- mean(teengamb_males$income)
avg_verbal <- mean(teengamb_males$verbal)


mean_gamble_males <- mean(teengamb_males$gamble)
n_males <- nrow(teengamb_males)
sd_gamble_males <- sd(teengamb_males$gamble)

# calclating our confidence intervale
se_gamble_males <- sd_gamble_males / sqrt(n_males)

ci_lower <- mean_gamble_males - qt(0.975, df = n_males - 1) * se_gamble_males
ci_upper <- mean_gamble_males + qt(0.975, df = n_males - 1) * se_gamble_males
```

```r
observed_ci <- data.frame(
  mean_gamble = mean_gamble_males,
  lcl = ci_lower,
  ucl = ci_upper
)

print(observed_ci)
```

```
##   mean_gamble      lcl      ucl
## 1      29.775 15.30219 44.24781
```

For this interval, it is based on actual observed values for males with average characteristics. The resulting interval is narrower than the model-based interval in part 1A. We have an estimate of 29.775 and an 95% confidence interval of (15.30, 44.25). This is expected because the interval in 1A is based on a prediction, which includes both data variability and model uncertainty, whereas the interval just calculated goes off of the data.

Question 2

We would want to consider the covariate of smoking on HIV and Stroke because smoking is a confounder. Additionally, we want to control for age because it is also a confounder. We would make sure to specify smoking and age in our model to account for these confounding variables so we can get more accurate estimations of the effect of HIV on stroke.