Homework 3

Elisabeth R Silver

2/11/2021

## Multiple Linear Regression with UT Austin Prof. Evaluation Data

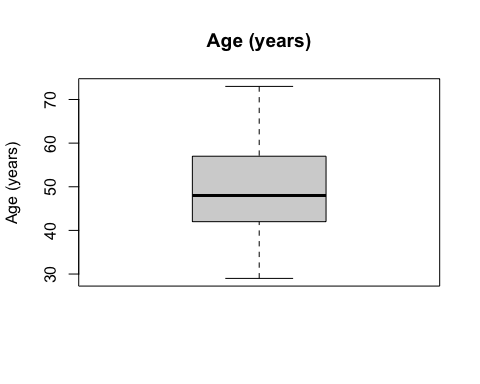
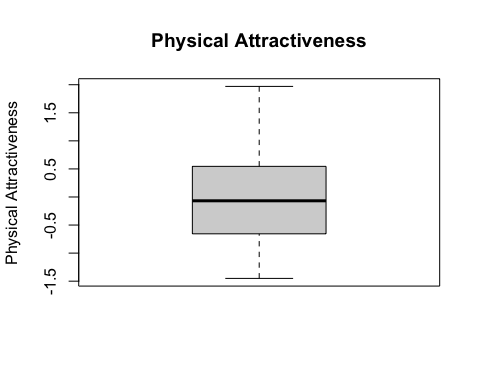
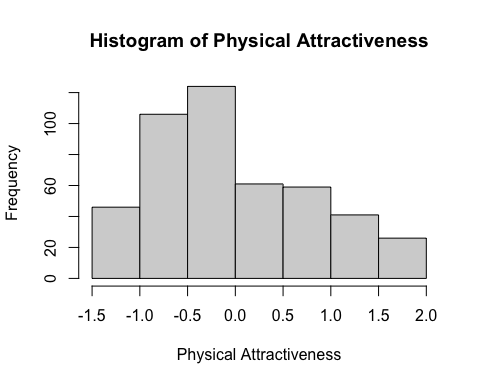
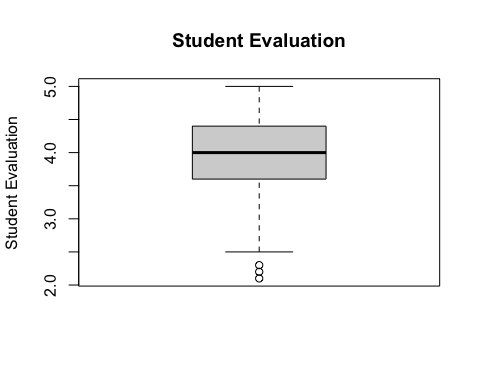
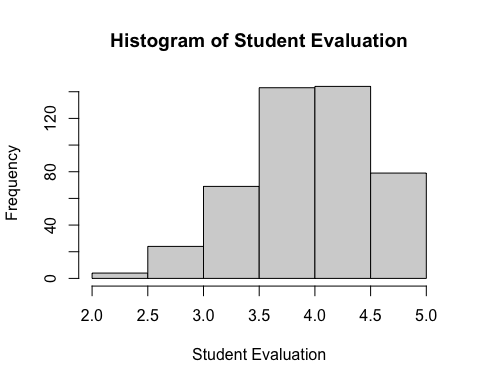
library(pacman)  
p\_load(dplyr, stats, psych, apaTables, ggplot2, ggpubr, AER, pwr, scales, papaja, knitr)  
data("TeachingRatings")  
  
vars <- c("eval", "beauty", "age")  
var.names <- c("Student Evaluation", "Physical Attractiveness", "Age (years)")

The dataset contains students’ course evaluations for 463 courses during the 2000-2002 academic years at UT Austin. It has (among other variables) data on the course instructor’s student evaluation, physical attractiveness (z-scored and rated by a panel of six students), and age.

### Variable exploration

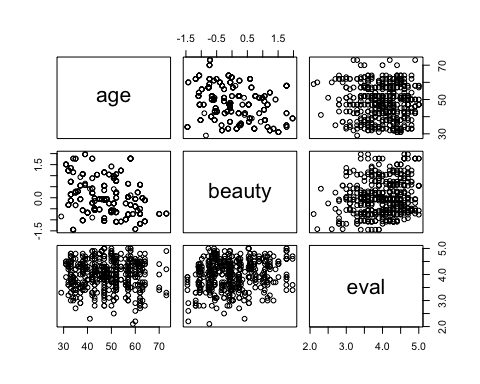
Student evaluations are negatively skewed with a few outliers on the lower end. Physical attractiveness is positively skewed with no apparent outliers. Age is more or less normally distributed with no apparent outliers. The mean for evaluation (*M* = 4.0, *SD* = 0.55) seems a bit high since the scale ranges from 0 to 5. Since attractiveness is already Z-scored, it’s not surprising that the mean is 0 (*SD* = 0.79). Similarly, the age is not surprisingly low or high (*M* = 48.37 years, *SD* = 9.80 years). UT Austin is a good school, so it’s plausible that instructors there have been in the business long enough to have established themselves (although there are some younger overachievers).

for(i in 1:length(vars)){  
 hist(TeachingRatings[[vars[i]]], xlab = var.names[i], main = paste0('Histogram of ', var.names[i]))  
 boxplot(TeachingRatings[[vars[i]]], main = var.names[i], ylab = var.names[i])  
}



Looking at pair plots, age and beauty seem closer to being multivariate normal than course evaluations and age and course evaluations and beauty. It seems like there are many observations that have the same course evaluations.

pairs(TeachingRatings[c(2,5,6)])



### Multiple Regression: Course evaluations regressed on instructor age and attractiveness

The model predicts that when an instructor’s physical attractiveness increases by one standard deviation (these are z-scored), their course evaluation would increaase by 0.13 on a 1-5 scale, holding age constant. This effect is significant, so the slope more than likely is not equal to 0.

Holding attractiveness constant, instructor age doesn’t have a strong relationship with course evaluations–a one year increase in age has almost zero association with change in evaluationn, controlling for attractiveness, and the p-value suggests as much.

The intercept indicates that when age and physical attractiveness equal 0. Since I don’t think a newborn baby would be qualified to teach a course (unless it’s in crying), the intercept isn’t too informative here.

The R2 seems pretty low, with age and attractiveness together explaining only 3.58% of the variance in course evaluations, which suggests that just being pretty and young really isn’t going to cut it in academia.

mod <- lm(eval~age+beauty, data = TeachingRatings)  
sum.mod <- summary(mod)  
summary(mod)

##   
## Call:  
## lm(formula = eval ~ age + beauty, data = TeachingRatings)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.80242 -0.36514 0.07407 0.39913 1.10206   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.9844013 0.1337296 29.794 < 2e-16 \*\*\*  
## age 0.0002868 0.0027148 0.106 0.916   
## beauty 0.1340634 0.0337441 3.973 8.24e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.546 on 460 degrees of freedom  
## Multiple R-squared: 0.03576, Adjusted R-squared: 0.03157   
## F-statistic: 8.53 on 2 and 460 DF, p-value: 0.0002305

### Power analysis

res1 <- pwr.f2.test(u = 2, v = 460, power = 0.8)  
res2 <- pwr.f2.test(u = 2, f2 = sum.mod$r.squared, power = 0.8)

Given the sample size, the maximum R2 possible with 80% power is 0.02. The minimum sample size required to obtain the observed R2 (0.04) is 272.

### APA-formatted outputs

tmp\_for\_table <- TeachingRatings %>% dplyr::select(c(eval, beauty, age))

#Dump the correlation into a separate doc. Annoyingly, APATables won't output using kable

apa.cor.table(tmp\_for\_table, filename = "homework 3 table 1.doc", table.number = 1, show.conf.interval = FALSE)

apa\_lm <- apa\_print(mod)  
apa\_lm$table['SE'] <- sum.mod$coefficients[,2]  
apa\_table(  
 apa\_lm$table,   
 caption = "Linear regression: Effect of Instructor Physical Attractiveness and Age on Course Evaluation"  
)

Table 1

*Means, standard deviations, and correlations with confidence intervals*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable | *M* | *SD* | 1 | 2 |
|  |  |  |  |  |
| 1. Evaluation | 4.00 | 0.55 |  |  |
|  |  |  |  |  |
| 2. Attractiveness | 0.00 | 0.79 | .19\*\* |  |
|  |  |  | [.10, .28] |  |
|  |  |  |  |  |
| 3. Age | 48.37 | 9.80 | -.05 | -.30\*\* |
|  |  |  | [-.14, .04] | [-.38, -.21] |
|  |  |  |  |  |

*Note.* *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). \* indicates *p* < .05. \*\* indicates *p* < .01.

Table 2

Linear regression: Effect of Instructor Physical Attractiveness and Age on Course Evaluation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Effect | Estimate | *SE* |  | 95% CI  *[LL, UL]* |  |
| Intercept | 3.98 | 0.13 |  | [3.72, 4.25] | < .001 |
| Age | 0.00 | 0.00 |  | [-0.01, 0.01] | .916 |
| Attractiveness | 0.13 | 0.03 |  | [0.07, 0.20] | < .001 |

*Note. SE* represents standard, *LL* and *UL* represent the lower and upper 95% confidence interval (*CI*) on the Estimate.