Simple Regression and Binary Predictors

Lab 4 Handout

Statistics 139

Topics

- Simple Regression Inference
- Regression diagnostics
- Categorical predictors with two levels
- Understanding R^2

Question 1: Simple linear regression The Prevention of REnal and Vascular END-stage Disease (PREVEND) study took place between 2003 and 2006 in the Netherlands. Clinical and demographic data for the 4,095 participants were collected. The data are provided to you in the `prevend.csv` data file.

As adults age, cognitive function changes, largely due to various cerebrovascular and neurodegenerative changes. The Ruff Figural Fluency Test (RFFT) is one measure of cognitive function; higher scores indicate better cognitive function. We will use linear regression to explore the relationship between age and RFFT score.

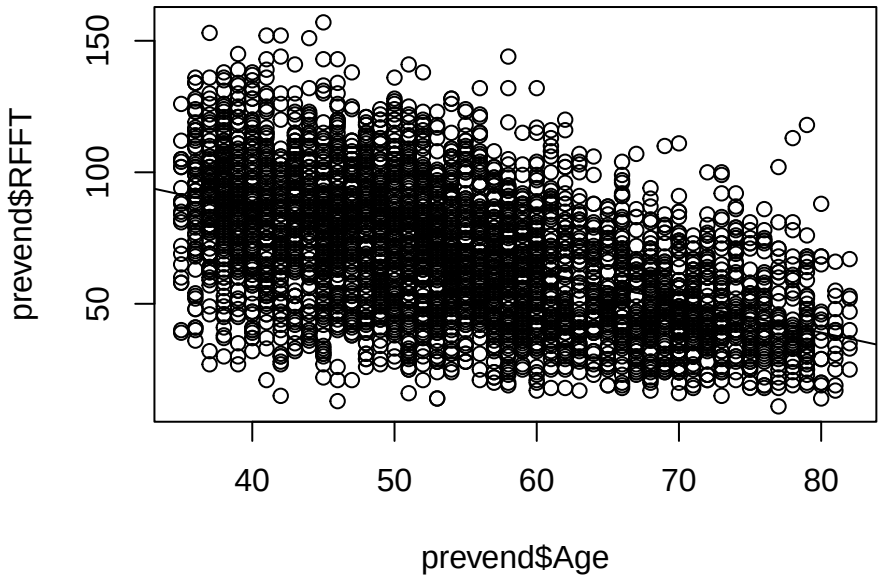
a) Create a scatterplot of RFFT score versus age and add a line of best fit.

```
prevend=read.csv("data/prevend.csv")

# create a plot, fit a model, and add the fitted line
model_1a <- lm(RFFT ~ Age, data=prevend)
summary(model_1a)
```

```
##
## Call:
## lm(formula = RFFT ~ Age, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.373 -15.636  -1.065  14.798  79.281
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  132.339      1.676   78.96   <2e-16 ***
## Age           -1.166      0.030  -38.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.26 on 4093 degrees of freedom
## Multiple R-squared:  0.2695, Adjusted R-squared:  0.2693
## F-statistic:  1510 on 1 and 4093 DF,  p-value: < 2.2e-16
```

```
plot(prevend$RFFT ~ prevend$Age)
abline(model_1a)
```



b) What are the slope and intercept values of the line of best fit? Interpret the values in the context of the data. Is the intercept value meaningful for this context?

Slope = 132.3394104, and the intercept = -1.1658715. The slope means that on average, when someone becomes one year older, their RFFT reduces by 1.17. The intercept means that when someone is 0 years old, their RFFT value on average is 132.34 (this isn't meaningful because extrapolation and babies aren't the smartest).

c) On average, how much does mean RFFT score differ between an individual who is 60 years old versus an individual who is 50 years old?

```
model_1a$coefficients[2]*(60-50)
```

```
##       Age
## -11.65872
```

d) Is it valid to use the linear model to estimate average RFFT score for an individual who is 20 years old? Explain your answer.

Not really because the min age is 35 in the dataset. This would be extrapolation.

```
summary(prevend$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   35.00   45.00   54.00   54.65   63.00   82.00
```

e) Formally test whether RFFT score is linearly associated with age. Calculate the associated 95% confidence interval and interpret the interval.

small p-value –> reject null hypothesis and conclude there's a linear association between RFFT score and age. the confidence interval for the slope is entirely negative, so the association is negative.

```
# the `summary` and `conf.int` commands on your `lm` object will be useful
summary(model_1a)
```

```
##
## Call:
## lm(formula = RFFT ~ Age, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.373 -15.636  -1.065  14.798  79.281
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  132.339      1.676   78.96   <2e-16 ***
## Age           -1.166      0.030  -38.86   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.26 on 4093 degrees of freedom
## Multiple R-squared:  0.2695, Adjusted R-squared:  0.2693
## F-statistic:  1510 on 1 and 4093 DF,  p-value: < 2.2e-16
```

```
confint(model_1a)
```

```
##                 2.5 %     97.5 %
## (Intercept) 129.05346 135.625361
## Age          -1.22469  -1.107053
```

f) Calculate a 95% confidence interval for the mean RFFT score of individuals who are 60 years old. Interpret the interval.

```
# the `predict` command on your `lm` object will be useful
newdata <- data.frame(Age = 60)
predict(model_1a, newdata = newdata, interval = "confidence", level=0.95)
```

```
##        fit      lwr      upr
## 1 62.38712 61.63615 63.13808
```

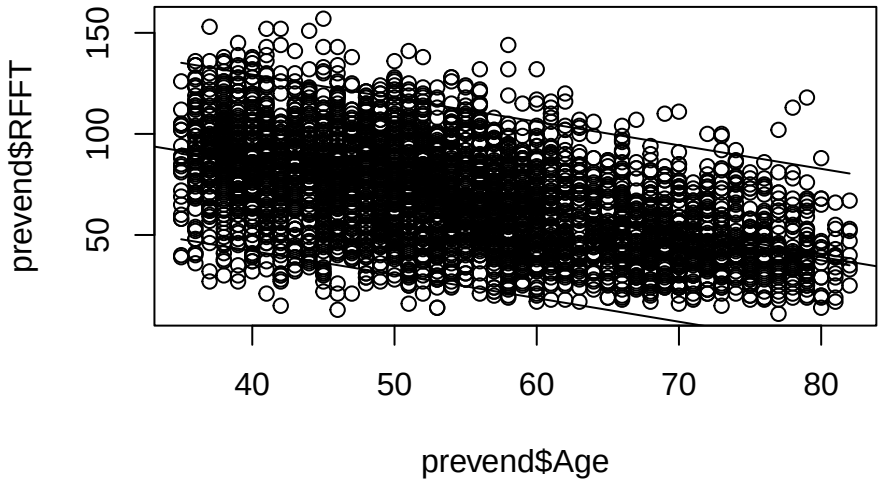
g) Calculate a 95% prediction interval for the RFFT score of a future individual (i.e., one who is not in the dataset) who is 60 years old. Interpret the interval.

```
# the `predict` command on your `lm` object will again be useful
predict(model_1a, newdata = newdata, interval = "prediction", level=0.95)
```

```
##        fit      lwr      upr
## 1 62.38712 18.74671 106.0275
```

h) Construct a plot of RFFT score versus age that shows the line of best fit, the 95% confidence bands, and the 95% prediction bands.

```r
#plot the data and line of best fit from before
plot(prevend$RFFT ~ prevend$Age)
abline(model_1a)

#plot confidence bands, you'll want to define a `dummy` x first
new_x <- seq(min(prevend$Age), max(prevend$Age))
conf.band <- predict(model_1a, newdata = data.frame(Age = new_x), interval = "confidence")
lines(new_x, conf.band[, 2])
lines(new_x, conf.band[, 3])

#plot prediction bands
pred_x <- seq(min(prevend$Age), max(prevend$Age))
conf.band <- predict(model_1a, newdata = data.frame(Age = pred_x), interval = "prediction")
lines(pred_x, conf.band[, 2])
lines(pred_x, conf.band[, 3])
```

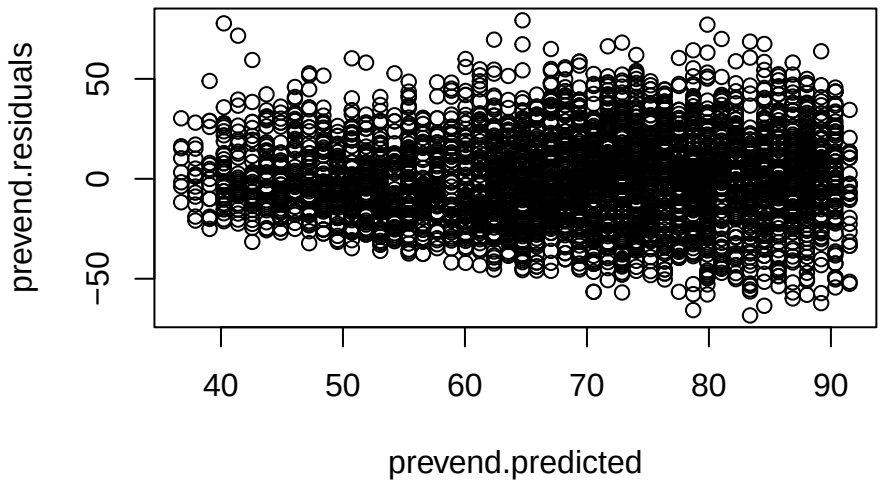


Question 2: Regression diagnostics There are five assumptions that must be met for a linear model to be considered reasonable: existence, linearity, independence, homogeneity of variance, and normally distributed residuals. Three of these can be checked.

Even though linearity and constant variability can be assessed from the scatterplot of y versus x, it is helpful to consult residual plots for a clearer view. Normality of residuals is best assessed through a normal probability plot; although skew can be visible from a histogram of the residuals, deviations from normality are more obvious on a normal probability plot.

a) Create a residual plot where the residual values (call them `prevend.residuals`) are plotted on the y-axis against predicted values (call them `prevend.predicted`) from the model on the x-axis, using data in `prevend`. Based on the plot, comment on the validity of the linearity and constant variability assumptions.

```
# residual plot.
prevend.residuals <- model_1a$residuals
prevend.predicted <- model_1a$fitted.values
plot(prevend.residuals ~ prevend.predicted)
```



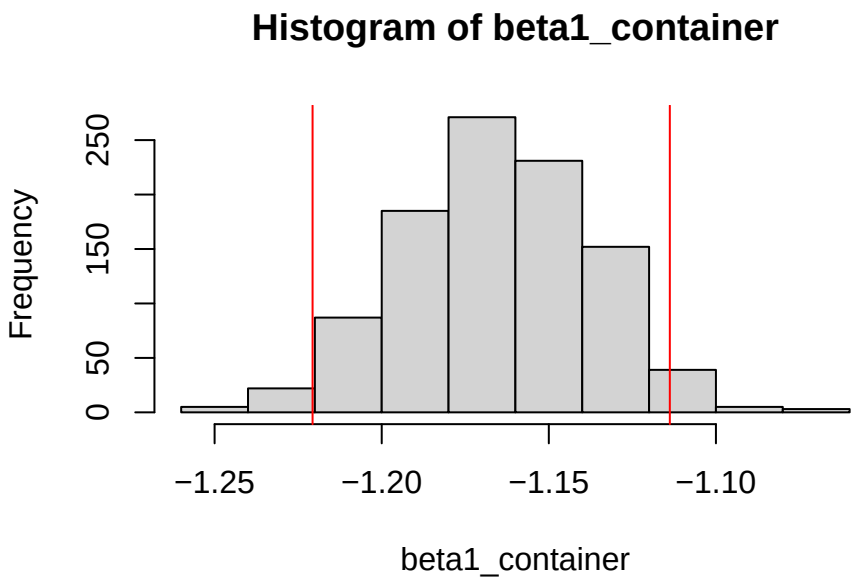
b) Run the code chunk below to create a normal probability plot of the residuals. For comparison purposes, the following figure shows a histogram of the residual values overlaid with a normal curve and the normal probability plot. Do the residuals seem approximately normally distributed?

c) Use a resampling approach to calculate a 95% confidence interval for β_1. Compare this interval to the one in problem 1.

```
# set parameters and seed
set.seed(139)
nboots <- 1000
n = nrow(prevend)
beta1_container = rep(NA, nboots)

# bootstrapping
for(i in 1:nboots){

  boot.index = sample(1:n, replace=T)
  boot.df = prevend[boot.index,]
  boot.lm = lm(RFFT~Age, boot.df)
  beta1_container[i] = coef(boot.lm)["Age"]

}

# determine the confidence interval
ci = quantile(beta1_container, c(0.025, 0.975))
ci
```

```
##      2.5%     97.5%
## -1.220682 -1.113793
```

```
# visualize distribution of sampling statistic with ci added
hist(beta1_container)
abline(v = ci, col="red")
```

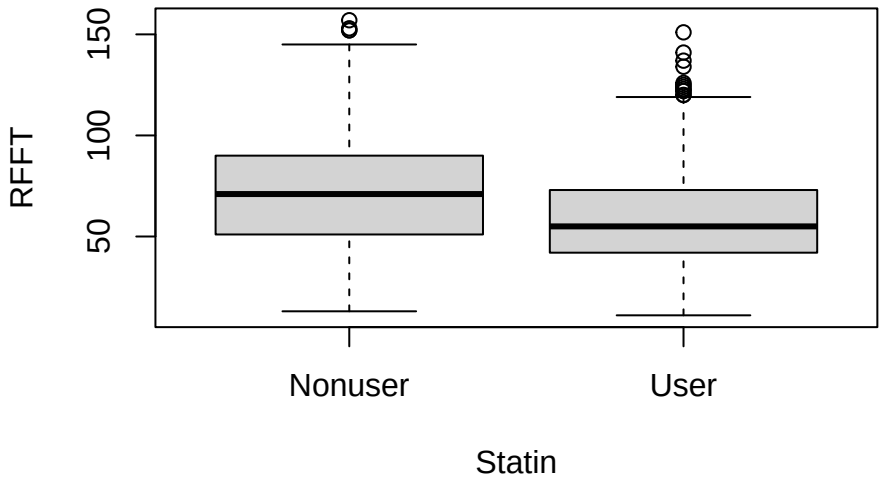

Histogram of beta1_container
Frequency
0
50
150
250
−1.25
−1.20
−1.15
−1.10
beta1_container

Question 3: Categorical predictors with two levels Statins are a class of drug widely used to lower cholesterol. However, some physicians have raised the question of whether treatment with a statin might be associated with an increased risk of cognitive decline.

Statin use is coded as a factor, where `Nonuser` represents a non-user and `User` represents a user.

a) Create a visual showing the association between RFFT score and statin use. Describe what you see. Calculate the mean RFFT score in each statin use group.

```
# plot
boxplot(RFFT~Statin, data=prevend)
```




```
# calculate means
tapply(prevend$RFFT, prevend$Statin, mean)
```

```
##  Nonuser     User
## 71.50799 58.44469
```

b) Fit a simple regression model. Calculate and interpret the slope coefficient, along with the associated 95% confidence interval.

```
# fit model
lm2 <- lm(RFFT ~ Statin, data=prevend)

# model summary
summary(lm2)
```

```
##
## Call:
## lm(formula = RFFT ~ Statin, data = prevend)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -58.508 -19.508  -1.445  17.492  92.555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.5080     0.4509  158.59   <2e-16 ***
## StatinUser  -13.0633     0.9596  -13.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.47 on 4093 degrees of freedom
## Multiple R-squared:  0.04331,    Adjusted R-squared:  0.04308
## F-statistic: 185.3 on 1 and 4093 DF,  p-value: < 2.2e-16
```

```
# confidence interval
confint(lm2)
```

```
##                 2.5 %    97.5 %
## (Intercept)  70.62401  72.39197
## StatinUser  -14.94472 -11.18188
```

c) Write the equation of the least-squares line and solve for the two possible values of $\widehat{RFFT}$. Confirm that the values match the ones from part (a).

```
lm2$coefficients[1]
```

```
## (Intercept)
##    71.50799
```

```
lm2$coefficients[1] + lm2$coefficients[2]
```

```
## (Intercept)
##    58.44469
```

d) Conduct a t-test for the difference in mean RFFT score between statin users and non-users. Compare the results of inference based on the linear model to those based on a two-group test.

```
#t-test
t.test(RFFT ~ Statin, data=prevend)
```

```
##
##  Welch Two Sample t-test
##
## data:  RFFT by Statin
## t = 14.774, df = 1652.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Nonuser and group User is not
## 95 percent confidence interval:
##  11.32901 14.79759
## sample estimates:
## mean in group Nonuser    mean in group User
##              71.50799              58.44469
```

```
t.test(RFFT ~ Statin, data=prevend, var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  RFFT by Statin
## t = 13.613, df = 4093, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Nonuser and group User is not
## 95 percent confidence interval:
##  11.18188 14.94472
## sample estimates:
## mean in group Nonuser    mean in group User
##              71.50799              58.44469
```

Question 4: Understanding R^2 The quantity R^2 describes the amount of variation in the response variable that is explained by the least squares line:

$$R^2 = \frac{\text{variance of predicted } y\text{-values}}{\text{variance of observed } y\text{-values}} = \frac{\text{Var}(\hat{y}_i)}{\text{Var}(y_i)}$$

R^2 can also be calculated using the following formula:

$$R^2 = \frac{\text{variance of observed } y\text{-values} - \text{variance of residuals}}{\text{variance of observed } y\text{-values}} = 1 - \frac{\text{Var}(e_i)}{\text{Var}(y_i)}$$

A simulation can be conducted in which y-values are sampled according to a population regression model $y = \beta_0 + \beta_1 x + \epsilon$, where the parameters β_0, β_1, and the standard deviation of ϵ are known. Recall that ϵ is a normally distributed error term with mean 0 and standard deviation σ.

a) Simulate 100 (x, y) values, where the values for x are 100 numbers randomly sampled from a standard normal distribution and the values for y are determined by the population model $y_i = 100 + 25x_i + \epsilon_i$, where $\epsilon_i \sim N(0, 5^2)$. Create a scatterplot of `y` versus `x` and add the line of best fit to the plot.

```
#set the seed

#simulate values

#plot the data with line of best fit
```

i. Does the line appear to be a good fit to the data?

ii. Why do the data points not fall exactly on a line, even though the data are simulated according to a known linear relationship between x and y?

iii. How well does the regression line estimate the population parameters β_0 and β_1?

```
#print model coefficients
```

iv. From a visual inspection, does it seem that the R^2 for this linear fit is relatively high or relatively low?

v. Using graphical summaries, compare the variances of the predicted and observed y-values; do they seem to have similar spread?

```
#plots
```

vi. Calculate the R^2 of the model.

```
#calculate R^2
```

b) Simulate 100 new (x, y) values. Like before, the x values are 100 numbers randomly sampled from a standard normal distribution and the y values are determined by the population model $y_i = 100 + 25x_i + \epsilon_i$. For these data, however, the error term is distributed $N(0, 50^2)$.

```
#clear the workspace
rm(list = ls())

#set the seed

#simulate values
```

i. Create a scatterplot of `y` versus `x` and add the line of best fit to the plot. Does the line appear to be a good fit to the data? How well does the regression line estimate the population parameters β_0 and β_1?

```
#plot the data with line of best fit
```

ii. Using graphical summaries, compare the variances of the predicted and observed y-values; do they seem to have similar spread?

```
#plots
```

iii. Based on the answers to parts (i) and (ii), do you expect the R^2 for this linear model to be relatively high or relatively low?

iv. Calculate the R^2 of the model.

```
#calculate R^2
```

c) Run the code chunk below to simulate 100 new (x, y) values based on a different true data-generative model.

```
#clear the workspace
rm(list = ls())

#set the seed
set.seed(2020)

#simulate values
n  = 100
x = rnorm(n)
error = rnorm(n, 0, 5)
y = 100 + 25*x + 5*x^2 + error
```

i. Fit a linear model predicting y from x to the data and calculate the R^2 for the model. Based on R^2, does the model seem to fit the data well?

```
#calculate R^2
```

ii. Plot the data and add the line of best fit. Evaluate whether the linear model is a good fit to the data; how does viewing the data change the conclusion from part (i)?

```
#plot the data
```