IEHC0046 BASIC STATISTICS FOR MEDICAL SCIENCES

Linear Regression: Practical I

11 November, 2020

In this practical session, we will be learning how to use R to fit and interpret a linear regression model. We will be using the dataset you have seen before in previous sessions and we will be examining the association between systolic blood pressure (sbp) and (continuous) age. We’ll also be using the tidyverse, summarytools, broom, multcomp packages. You will need to install multcomp and broom.

Note, please load the packages in the order below. multcomp loads a function called select(). tidyverse also contains a function with this name (which we want to use) - loading tidyverse later will mean that select() from that package is available.

# install.packages(c("multcomp", "broom")) # Uncomment and  
library(multcomp)  
library(tidyverse)  
library(broom)  
library(summarytools)  
load("elsa.Rdata")

There are three sections to this practical:

1. Visualising relationships between variables
2. Linear regression against a single continuous variable (age)
3. Testing the assumptions of linear regression

# 1. Visualising relationships between variables

## 1.1. Summary statistics

We are going to examine the relationship between systolic blood pressure (our outcome - or dependent - variable) sbp and age in years (our exposure - or independent - variable). We first need to explore these variables in our dataset.

descr(elsa$age)

## Descriptive Statistics   
## elsa$age   
## Label: Age in years   
## N: 3129   
##   
## age  
## ----------------- ---------  
## Mean 60.12  
## Std.Dev 9.60  
## Min 45.00  
## Q1 52.00  
## Median 59.00  
## Q3 67.00  
## Max 90.00  
## MAD 11.86  
## IQR 15.00  
## CV 0.16  
## Skewness 0.45  
## SE.Skewness 0.04  
## Kurtosis -0.71  
## N.Valid 3129.00  
## Pct.Valid 100.00

**Q: What is the mean age of our sample? What is the median age?**

We can see here that the mean age is 60.1 years and the median age (50th percentile) is 59 years. Age has been censored (collapsed) at 45 and 90 years.

descr(elsa$sbp)

## Descriptive Statistics   
## elsa$sbp   
## Label: Systolic BP in mmHg   
## N: 3129   
##   
## sbp  
## ----------------- ---------  
## Mean 140.38  
## Std.Dev 19.65  
## Min 88.50  
## Q1 126.00  
## Median 138.50  
## Q3 151.50  
## Max 222.50  
## MAD 18.53  
## IQR 25.50  
## CV 0.14  
## Skewness 0.68  
## SE.Skewness 0.05  
## Kurtosis 0.66  
## N.Valid 2692.00  
## Pct.Valid 86.03

**Q: What is the mean and median systolic blood pressure for our sample?**

We can see here that the mean systolic blood pressure is 140.4 mmHg (millimeters of mercury) and the median systolic blood pressure is 138.5 mmHg (50th percentile). The mean is slightly higher than median, suggesting that systolic blood pressure has a slightly positively skewed distribution in our sample.

**Q: How many observations do we have for age and for systolic blood pressure?**

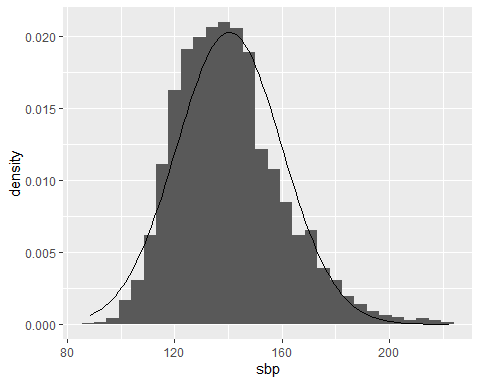
We have 3,129 observations for age and 2,692 for systolic blood pressure. This means all participants answered the question about their age at last birthday or date of birth and so we have no missing observations. However, we do have missing values for systolic blood pressure (coded as NA in the dataset). This could be for a number of reasons, for instance, fewer people consented to have their blood pressure measurements taken or it not being possible for the nurse to take an accurate blood pressure measurement. We are going to just run the analysis on those participants with valid systolic blood pressure measurements for now (R will delete them from the analysis automatically as they are coded as NA). We would need to be sure that we document this in our methods section.

A histogram is be a useful graph to use here to see how our outcome (systolic blood pressure) is distributed and to assess whether it approximates a normal distribution type. We can use ggplot (part of the tidyverse) to draw a histogram for the variable sbp and overlay a normal distribution for comparison. We saw the code to do this in a previous practical.

ggplot(elsa) +  
 aes(x = sbp) +  
 geom\_histogram(aes(y = ..density..)) +  
 stat\_function(fun = dnorm, args = list(mean = mean(elsa$sbp, na.rm = TRUE),  
 sd = sd(elsa$sbp, na.rm = TRUE)))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 437 rows containing non-finite values (stat\_bin).



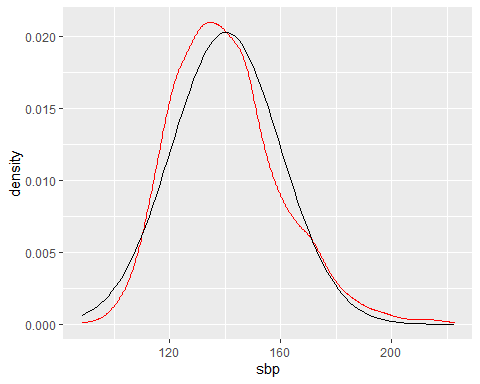
**Q: Is systolic blood pressure normally distributed?**

Almost, but it is slightly positively skewed (the tail on the right side of the distribution is slightly longer).

Histograms are sensitive to the number and positioning of bins (or columns) that are used in the plot. An alternative to the histogram is the kernel density plot, which approximates the probability density of a variable. Kernel density plots have the advantage of being smooth, unlike histograms. The relevant geom\_\*() function is geom\_density(). We don’t need to specify that the derived variable ..density.. is mapped to the y-axis in this case, because it is the default mapping for this function. Let’s change the colour of the density function, though, so we can distingish it from the normal curve.

ggplot(elsa) +  
 aes(x = sbp) +  
 geom\_density(color = "red") +  
 stat\_function(fun = dnorm, args = list(mean = mean(elsa$sbp, na.rm = TRUE),  
 sd = sd(elsa$sbp, na.rm = TRUE)))

## Warning: Removed 437 rows containing non-finite values (stat\_density).



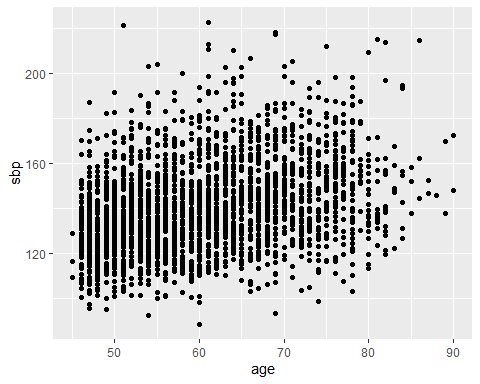
The plot confirms that the distribution of systolic blood pressure is slightly positively skewed.

## 1.2. Association between two continuous variables

To examine the relationship between systolic blood pressure and age we can start by plotting a scatter plot. This graph will plot systolic blood pressure (y axis) against age (x axis) for each observation in our sample with valid values of sbp and age. The relevant geom\_\*() function from ggplot is geom\_point().

ggplot(elsa) +  
 aes(x = age, y = sbp) +  
 geom\_point()

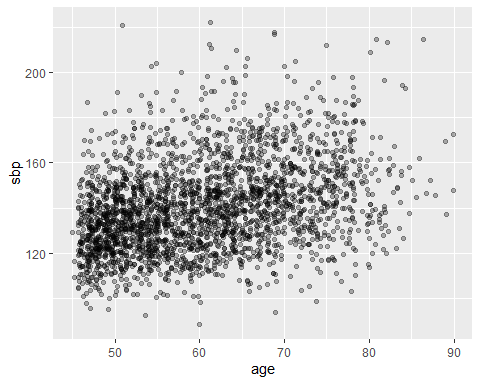
## Warning: Removed 437 rows containing missing values (geom\_point).



An issue with this plot is that many of the points are overlapping (because age is measured as in integer years). We can use the geom\_jitter() function to add a bit of random “noise” to the data so the points don’t overlap. We can also change the transparency (using the argument alpha) so a mass of overlapping points can be distinguished. alpha can be set between 0 (transparent) and 1 (solid colour).

ggplot(elsa) +  
 aes(x = age, y = sbp) +  
 geom\_jitter(alpha = 0.3)

## Warning: Removed 437 rows containing missing values (geom\_point).



**Q: What relationship exists between systolic blood pressure and age?**

From the scatter plot, higher values of age in general correspond with higher values of blood pressure, so there is a positive linear relationship suggested, although this is quite weak. We know that systolic blood pressure usually increases with age and we see this here. The linear relationship is not very steep in comparison with the wide spread of blood pressure values for any particular age - knowing someone’s age might help us predict sbp a little better than just using the sample mean, but our guess is still likely to be off!

We can also get a numerical estimate of how age and systolic blood pressure are related by running a correlation.

elsa %>%  
 select(age, sbp) %>%  
 cor(use = "complete.obs")

## age sbp  
## age 1.0000000 0.3396941  
## sbp 0.3396941 1.0000000

**Q: What is the sign of the correlation? Is that what you expect from the scatter diagram? What about its magnitude?**

The sign is positive, which is what we would expect as we have already observed that there is a weak but positive linear relationship between age and blood pressure in our data. Its magnitude is 0.34 which is less than halfway between 0 (i.e. no relationship) and 1 (perfect relationship). This also makes sense as the relationship is certainly not perfect in our data.

# 2. Linear regression against a single continuous variable (age)

We will assume that the association between systolic blood pressure and age is linear and fit a linear regression model:

This equation is a way of representing the linear regression model we will be estimating. is systolic blood pressure for ELSA participant *i* and is their age when their blood pressure was measured. The difference between the actual systolic blood pressure for participant i and that predicted in the model is the residual, .

## 2.1. Linear regression in R

To estimate a linear regression in R you use the lm() function. The linear model is represented with the formula notation: y ~ x. To view a nice summary of the linear regression object use summary().

mod <- lm(sbp ~ age, elsa)  
summary(mod)

##   
## Call:  
## lm(formula = sbp ~ age, data = elsa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -52.92 -12.87 -1.60 10.56 87.01   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 98.76273 2.25024 43.89 <2e-16 \*\*\*  
## age 0.69076 0.03688 18.73 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.48 on 2690 degrees of freedom  
## (437 observations deleted due to missingness)  
## Multiple R-squared: 0.1154, Adjusted R-squared: 0.1151   
## F-statistic: 350.9 on 1 and 2690 DF, p-value: < 2.2e-16

**Q: What is the value of R-squared, and what does it mean?**

which means that only 12% of the variability of systolic blood pressure is explained by age. This statistic is also the square of the correlation coefficient that we saw earlier. This reflects what we saw from the scatter plot, that there was a large spread of blood pressure values around each age point and that the linear relationship was not very steep. It is not surprising then that this model, which includes just age, explains a relatively small proportion of the variability in systolic blood pressure.

**Q: What is the estimated slope coefficient for age, and what does it mean?**

The estimated coefficient for age is 0.69, which we interpret as the absolute increase in systolic blood pressure predicted as the age of participants increases by a single unit. Since age is measured in years, this means that per 1-year increase in age there is a 0.69 mmHg (millimeters of mercury) increase in systolic blood pressure.

**Q: What is the coefficient “(Intercept)” and what can it tell us?**

In a linear regression, the coefficient for “(Intercept)” is the estimated intercept, which here is 98.76. This is the predicted value of sbp when age equals zero. In other words, this is the in the linear equation. We should note that given, the minimum age in our dataset is 45, this is an extrapolation beyond the data we use to estimate the model. It is unlikely to be correct in practice, and we should be wary of extrapolating beyong the model without further justification.

**Q: What do the p-values tell us?**

The p-value is from a t-test with a null hypothesis that the true coefficent is equal to 0. Both p-values are less than 0.01 so this suggests that there is a (positive) relationship between age and sbp and that sbp where age is zero is not zero. Remember, though, that this latter result is an extrapolation.

We might also want to see the confidence intervals for the model coefficients. We can get these with the confint() function.

confint(mod)

## 2.5 % 97.5 %  
## (Intercept) 94.3503613 103.1750933  
## age 0.6184553 0.7630705

Neither of these overlap zero, which is to be expected given the p-values were both less than 0.05.

## 2.2. Predicted values from a regression model

We can use these model results to make predictions about systolic blood pressure using the fitted regression line. Although as we will see in later sessions, before we go on to use these predicted values in our analysis, we will need to consider how certain we are that the line fits our data.

An assumption of the linear regression model is that the residual errors are equal to 0, on average, and that residual errors are uncorrelated with the independent variables. Therefore our best prediction of a person’s sbp if we know their age is:

We estimated and above, so we can plug them into this equation.

**Q: What is the systolic blood pressure for someone aged 55?**

We can work this out by hand.

98.8 + 0.69\*55

## [1] 136.75

The predicted systolic blood pressure for someone aged 55 is 136.75 mmHg.

Alternatively, we can use coefficients directly using the coef() function. This returns a named vector of the coefficients.

coef(mod)

## (Intercept) age   
## 98.7627273 0.6907629

coef(mod)["(Intercept)"] + coef(mod)["age"]\*55

## (Intercept)   
## 136.7547

(We can ignore the fact that the results has the name “(Intercept)” - this has just occurred because we subsetted the vector).

A final approach is to use the glht() function from the package multcomp. To use glht() pass the function a model and a string giving the combination of variables you want to test. confint() displays the confidence intervals for the glht() result.

glht(mod, "`(Intercept)` + 55\*`age` = 0") %>%  
 confint()

##   
## Simultaneous Confidence Intervals  
##   
## Fit: lm(formula = sbp ~ age, data = elsa)  
##   
## Quantile = 1.9608  
## 95% family-wise confidence level  
##   
##   
## Linear Hypotheses:  
## Estimate lwr upr   
## `(Intercept)` + 55 \* age == 0 136.7547 135.9596 137.5498

We can also ask R to calculate a predicted value of systolic blood pressure for everyone in our sample using the function augment() from the package broom. augment() produces a new data from with the model variables (age and sbp) and some extra variables based on the results of the model. The predicted value is stored in a variable .fitted. We can use select() and slice() to see the first five values.

aug <- augment(mod)  
aug %>%  
 select(sbp, age, .fitted) %>%  
 slice(1:5)

## # A tibble: 5 x 3  
## sbp age .fitted  
## <dbl> <dbl> <dbl>  
## 1 145 51 134.  
## 2 119 67 145.  
## 3 132. 47 131.  
## 4 164 74 150.  
## 5 139 47 131.

augment() also stores the residuals in a variable .resid.

aug %>%  
 select(sbp, age, .fitted, .resid) %>%  
 slice(1:5)

## # A tibble: 5 x 4  
## sbp age .fitted .resid  
## <dbl> <dbl> <dbl> <dbl>  
## 1 145 51 134. 11.0   
## 2 119 67 145. -26.0   
## 3 132. 47 131. 0.271  
## 4 164 74 150. 14.1   
## 5 139 47 131. 7.77

We can see that the residual is equal to the observed outcome (sbp) minus the predicted values (.fitted).

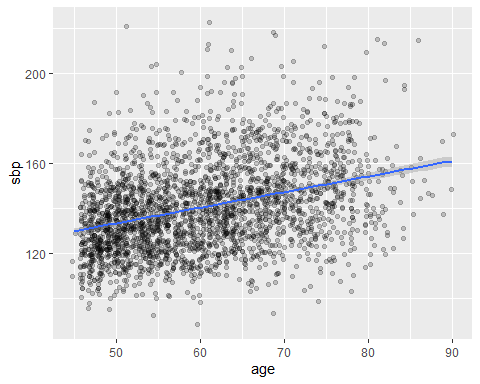
It can be also useful to see a scatterplot of each pair of age and systolic blood pressure values together with the fitted regression line (the predicted values). To do this use we can add the geom\_smooth() function to a ggplot function call. geom\_smooth() produces a regression line behind the scenes. We want a linear regression, so we need to specify the function lm() within geom\_smooth().

ggplot(elsa) +  
 aes(x = age, y = sbp) +  
 geom\_jitter(alpha = 0.2) +  
 geom\_smooth(method = lm)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 437 rows containing non-finite values (stat\_smooth).

## Warning: Removed 437 rows containing missing values (geom\_point).



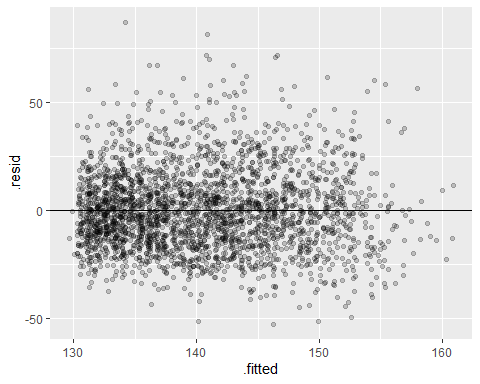
# 3. Testing the assumptions of linear regression

We need to check the assumptions of the regression model that were introduced in the lecture.

## 3.1. Plots of residuals against fitted values

In a well-fitted model, there should be no pattern to the residuals plotted against the predicted values. We can enter the data frame produced by augment() function into ggplot() to do this. Below, we use geom\_hline() to drop a horizontal line running through - remember, residuals should be equal to 0 on average and have constant variance (homoskedasticity).

ggplot(aug) +  
 aes(x = .fitted, y = .resid) +  
 geom\_hline(yintercept = 0) +  
 geom\_jitter(alpha = 0.2)



**Q: Is there evidence of heteroskedasticity (the opposite of homoskedasticity)?**

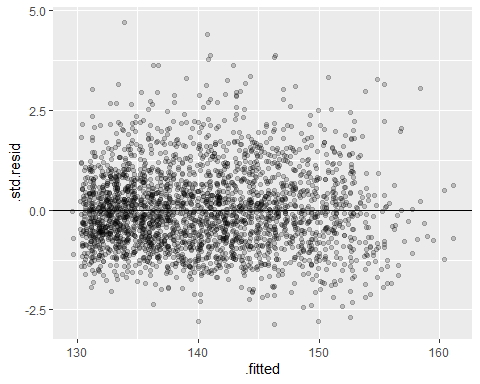
No, there is no heteroskedasticity, as there is no increasing or decreasing variation in the residuals as the fitted values increase.

**Q: Are the residuals normally distributed?**

Almost, although the pattern of residuals is slightly positively skewed, as there are some residuals at the top of the graph (absolute values >50).

You can also look at the graph of standardized residuals (i.e., residual divided by its standard deviation) versus fitted values. augment() produces standardized residuals (variable .std.resid). If the residuals were normally distributed, the standardized residuals would be distributed as a standard normal variable (i.e. with a mean of 0 and a standard deviation of 1) and you would expect 95% of them to be between -1.96 and +1.96.

ggplot(aug) +  
 aes(x = .fitted, y = .std.resid) +  
 geom\_hline(yintercept = 0) +  
 geom\_jitter(alpha = 0.2)



Similar to what we found before, standardized residuals are almost normally distributed, although the pattern of standardized residuals is slightly positively skewed, as there are more residual scatters at the top of graph (between 2.5 and 5).

## 3.2. Checking for normally distributed standardized residuals visually

To check the normality assumption more precisely, we can produce a kernel density plot again, overlaying a normal distribution (mean = 0, SD = 1) on top. In this case, we don’t need to pass any other arguments to stat\_function() because, by default, dnorm() produces normal distributions with mean = 0 and standard deviation = 1.

ggplot(aug) +  
 aes(x = .std.resid) +  
 geom\_density(color = "red") +  
 stat\_function(fun = dnorm)



The kernel density plot confirms that the distribution of standardized residuals is almost normally distributed, although is slightly positively skewed.