IEHC0046 BASIC STATISTICS FOR MEDICAL SCIENCES

Linear Regression: Practical II

11 November, 2020

In this practical session, we will be learning how to use R to fit and interpret a linear regression model against a binary or a categorical variable. We will also run linear regression models with multiple predictor variables. 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 age, sex, social class, wealth and smoking. We’ll also be using the tidyverse, summarytools, multcomp and car packages. You will need to install the car package.

# install.packages("car") # Uncomment to install  
library(car)  
library(multcomp)  
library(tidyverse)  
library(summarytools)  
load("elsa.Rdata")

The four sections to this practical are as follows:

1. Linear regression against a binary variable
2. Linear regression against a categorical variable
3. Multiple regression
4. Optional exercise

# 1. Linear regression against a binary variable

# 1.1. Examining and visualising variables

We will use the freq() function to tabulate our variable sex. We’ll then use stby() with descr() to see the distribution of sbp by sex.

freq(elsa$sex)

## Frequencies   
## elsa$sex   
## Label: Sex   
## Type: Factor   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## ------------ ------ --------- -------------- --------- --------------  
## male 1368 43.72 43.72 43.72 43.72  
## female 1761 56.28 100.00 56.28 100.00  
## <NA> 0 0.00 100.00  
## Total 3129 100.00 100.00 100.00 100.00

stby(elsa$sbp, elsa$sex, descr)

## Descriptive Statistics   
## sbp by sex   
## Data Frame: elsa   
## Label: Systolic BP in mmHg   
## N: 1368   
##   
## male female  
## ----------------- --------- ---------  
## Mean 141.20 139.74  
## Std.Dev 17.94 20.87  
## Min 88.50 92.50  
## Q1 128.50 124.00  
## Median 139.00 137.50  
## Q3 151.50 151.25  
## Max 217.00 222.50  
## MAD 17.05 20.02  
## IQR 23.00 27.12  
## CV 0.13 0.15  
## Skewness 0.65 0.71  
## SE.Skewness 0.07 0.06  
## Kurtosis 0.80 0.53  
## N.Valid 1184.00 1508.00  
## Pct.Valid 86.55 85.63

**Q: What percentage of the sample are men and what is the mean systolic blood pressure for men and for women in the sample?**

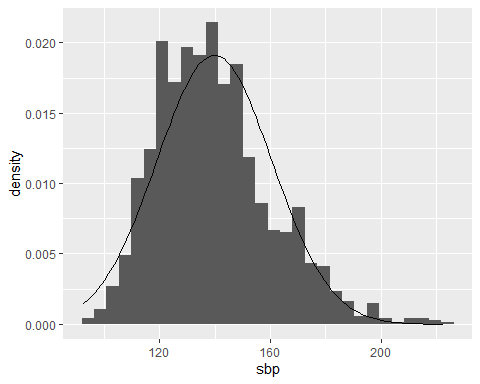
We can see that around 44% of our sample are men (n=1,368). Of these, 1,184 men have a valid systolic blood pressure measurement and the mean systolic blood pressure for men in this sample is 141.2 mmHg. Compared to this, 1,508 women have a valid systolic blood pressure and the mean for women in our sample is 139.7 mmHg, so mean systolic blood pressure is slightly lower in women.

We can also examine a histogram for sbp separately for men and women. We can see that both are a bit positively skewed, but approximate a normal distribution.

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

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

## Warning: Removed 253 rows containing non-finite values (stat\_bin).



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

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

## Warning: Removed 184 rows containing non-finite values (stat\_bin).



## 1.2. Linear regression

We will run the linear regression using the variable sex. The lm() function call has the same syntax with a binary independent variable and a continuous one. The linear equation with sex as the exposure is:

So that the model can be estimated, in the model sex = male will be inputted numerically by R as sex = 0; sex = female will be inputted as sex = 1.

mod\_1 <- lm(sbp ~ sex, elsa)  
summary(mod\_1)

##   
## Call:  
## lm(formula = sbp ~ sex, data = elsa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -52.698 -14.243 -2.198 10.802 82.757   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 141.1976 0.5707 247.390 <2e-16 \*\*\*  
## sexfemale -1.4543 0.7626 -1.907 0.0566 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.64 on 2690 degrees of freedom  
## (437 observations deleted due to missingness)  
## Multiple R-squared: 0.00135, Adjusted R-squared: 0.0009789   
## F-statistic: 3.637 on 1 and 2690 DF, p-value: 0.05662

**Q: What is the value of the coefficient for female? What does it mean?**

When a factor is used as an independent variable in R, the resulting coefficient is named as var\_namecategory, so the coefficient for female is sexfemale. The value of the coefficient for female is -1.45, which means that compared to men, women have lower blood pressure by 1.45 mmHg. We can also think of this as the predicted change in the outcome when the exposure is one instead of zero. Again, we should also examine the 95% confidence intervals, which show us the possible lower and upper value for our coefficient.

confint(mod\_1)

## 2.5 % 97.5 %  
## (Intercept) 140.078483 142.31678679  
## sexfemale -2.949558 0.04102484

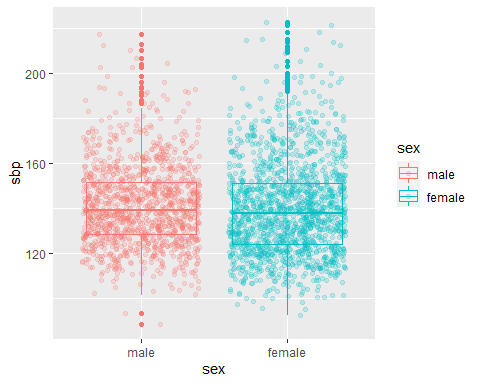
The 95% CI for difference in blood pressure for men and women spans -2.95-0.04. These two values give 95% coverage of the population coefficients that are compatible with the study data. The t-test for sex tests H0 (the null hypothesis) that there is no relationship between sex and systolic blood pressure. The p value associated with this test is relatively low (p=0.057) so there is some statistical evidence against the null hypothesis.

**Q: What is the value of the R2 and what is its meaning?** The R-squared is now 0.0014, which means that sex only explains 0.14% of the variability in systolic blood pressure! So there is a lot of unexplained variability in this model. We can see this by making a comparing boxplots of the distributions of sbp by sex. Let’s also add a layer of points (geom\_jitter) to see the underlying values. The two distributions are very similar. Our ability to predict sbp is not much helped by knowing the person’s sex.

ggplot(elsa) +  
 aes(x = sex, y = sbp, color = sex) +  
 geom\_jitter(alpha = 0.2) +  
 geom\_boxplot(fill = NA)

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

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



## 1.3. Predicted values from a regression model

**Q. What is the predicted systolic blood pressure for men?**

Recall, that 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 they are male is:

In other words, the predicted blood pressure for men is or “(Intercept)” in the output above

coef(mod\_1)["(Intercept)"] + coef(mod\_1)["sexfemale"]\*0

## (Intercept)   
## 141.1976

**Q: What is the predicted systolic blood pressure for women?** Our best prediction of a person’s sbp if we know they are female is:

coef(mod\_1)["(Intercept)"] + coef(mod\_1)["sexfemale"]\*1

## (Intercept)   
## 139.7434

Notice that these values match the mean sbp values for men and women we calculated earlier and the coefficient for female is the difference between them. Linear regression against a single binary variable gives you the same result of a t-test.

t.test(sbp ~ sex, elsa)

##   
## Welch Two Sample t-test  
##   
## data: sbp by sex  
## t = 1.9418, df = 2667.9, p-value = 0.05227  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.01426405 2.92279692  
## sample estimates:  
## mean in group male mean in group female   
## 141.1976 139.7434

Tip: You can also use R to create predicted values and residuals as we did in the last practical session. The scatter plot would be just 2 straight lines: one for men and one for women. There’s no need to check for satisfying the normality assumptions when you only have binary independent variables or categorical independent variables in a linear regression model.

**Extra: What do you think will happen to the intercept and the direction of the coefficient if you run the regression model using a new variable which reverses the categories of sex (use fct\_rev() from forcats to do this)**

# 2. Linear regression with a single categorical variable (social class)

Next, we are going to run a regression model using a single categorical variable and we will be examining the association between systolic blood pressure and social class (sclass).

## 2.1. Examing and visualising variables

**Q: How many categories of social class are in our sample? Should we recode this variable?**

freq(elsa$sclass)

## Frequencies   
## elsa$sclass   
## Label: Social Class   
## Type: Factor   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## --------------------------------- ------ --------- -------------- --------- --------------  
## I - Professional 155 5.05 5.05 4.95 4.95  
## II - Managerial technical 923 30.05 35.09 29.50 34.45  
## IIIN - Skilled non-manual 766 24.93 60.03 24.48 58.93  
## IIIM - Skilled manual 554 18.03 78.06 17.71 76.64  
## IV - Semi-skilled manual 460 14.97 93.03 14.70 91.34  
## V - Unskilled manual 202 6.58 99.61 6.46 97.79  
## Armed forces 4 0.13 99.74 0.13 97.92  
## Not fully described 8 0.26 100.00 0.26 98.18  
## <NA> 57 1.82 100.00  
## Total 3129 100.00 100.00 100.00 100.00

We have some missing values. We are going to just run the analysis on those participants with valid measurements for now. We can see that there are eight categories in our social class variable, however, two of these categories are very small and less useful (armed forces, not fully described) and there are also some very small numbers in some of the other categories. We will recode this variable into three groups. There are pros and cons to doing this and it is important to think about what this might mean for your research. This social class classification used here is The Registrar General’s Social Class Scale (1911).

We will use the case\_when() function to create a new variable sclass\_3 which collapses these categories. (as.numeric(sclass) converts the factor to the underlying integer vector which makes the case\_when() statements easier to write!)

elsa <- elsa %>%  
 mutate(sclass\_n = as.numeric(sclass),  
 sclass\_3 = case\_when(sclass\_n %in% 1:2 ~ "Prof/Managerial",  
 sclass\_n == 3 ~ "Skilled Non-Manual",  
 sclass\_n %in% 4:6 ~ "Manual/Routine") %>%  
 factor(c("Prof/Managerial", "Skilled Non-Manual",   
 "Manual/Routine")))  
table(elsa$sclass, elsa$sclass\_3, useNA = "ifany")

##   
## Prof/Managerial Skilled Non-Manual Manual/Routine  
## I - Professional 155 0 0  
## II - Managerial technical 923 0 0  
## IIIN - Skilled non-manual 0 766 0  
## IIIM - Skilled manual 0 0 554  
## IV - Semi-skilled manual 0 0 460  
## V - Unskilled manual 0 0 202  
## Armed forces 0 0 0  
## Not fully described 0 0 0  
## <NA> 0 0 0  
##   
## <NA>  
## I - Professional 0  
## II - Managerial technical 0  
## IIIN - Skilled non-manual 0  
## IIIM - Skilled manual 0  
## IV - Semi-skilled manual 0  
## V - Unskilled manual 0  
## Armed forces 4  
## Not fully described 8  
## <NA> 57

Now we can look at the mean systolic blood pressure for each of the different social class groups.

stby(elsa$sbp, elsa$sclass\_3, descr)

## Descriptive Statistics   
## sbp by sclass\_3   
## Data Frame: elsa   
## Label: Systolic BP in mmHg   
## N: 1078   
##   
## Prof/Managerial Skilled Non-Manual Manual/Routine  
## ----------------- ----------------- -------------------- ----------------  
## Mean 138.55 139.97 141.78  
## Std.Dev 18.80 19.45 20.10  
## Min 88.50 97.50 95.00  
## Q1 125.00 126.00 127.50  
## Median 136.00 138.75 140.00  
## Q3 149.50 150.50 152.75  
## Max 222.50 215.00 218.00  
## MAD 17.79 18.90 18.53  
## IQR 24.50 24.38 25.12  
## CV 0.14 0.14 0.14  
## Skewness 0.62 0.62 0.68  
## SE.Skewness 0.08 0.10 0.08  
## Kurtosis 0.56 0.39 0.65  
## N.Valid 937.00 658.00 1040.00  
## Pct.Valid 86.92 85.90 85.53

**Q: What is the mean systolic blood pressure for the three groups?**  Professional/Managerial (138.6); Skilled non-manual (140.0); Manual/Routine (141.8). These seem to suggest a social gradient in blood pressure levels. We will check this using a linear regression.

## 2.2. Linear Regression

We will now perform a linear regression of sbp against sclass\_3g. This will allow us to make comparisons of the outcome between the three social class groups.

Because sclass\_3 contains more than two categories, we actually specify the linear model with dummy variables (where is the number of categories): one dummy variable (0/1) for whether the participant is skilled non-manual, another for whether they are manual/routine, with the reference category as professional/managerial. Note these dummy variables are mutually exclusive - a participant can have the dummy = 1 for at most one dummy variable.

mod\_2 <- lm(sbp ~ sclass\_3, elsa)  
summary(mod\_2)

##   
## Call:  
## lm(formula = sbp ~ sclass\_3, data = elsa)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -50.053 -13.972 -1.777 10.947 83.947   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 138.5534 0.6364 217.697 < 2e-16 \*\*\*  
## sclass\_3Skilled Non-Manual 1.4185 0.9909 1.432 0.152393   
## sclass\_3Manual/Routine 3.2240 0.8775 3.674 0.000243 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.48 on 2632 degrees of freedom  
## (494 observations deleted due to missingness)  
## Multiple R-squared: 0.00514, Adjusted R-squared: 0.004384   
## F-statistic: 6.8 on 2 and 2632 DF, p-value: 0.001134

confint(mod\_2)

## 2.5 % 97.5 %  
## (Intercept) 137.3053700 139.801354  
## sclass\_3Skilled Non-Manual -0.5245056 3.361551  
## sclass\_3Manual/Routine 1.5033689 4.944715

**Q: How is the variable social class related to systolic blood pressure?**

The coefficient for skilled non-manual is 1.42 meaning that workers in these occupations have 1.42 mmHg higher estimated blood pressure than professional/managerial workers. However, the confidence intervals for this coefficient are quite wide (-0.52, 3.36, perhaps because this is quite a small category. The p value associated with this test is 0.152, so there is little statistical evidence against the null hypothesis. Remember, this does not mean that there is actually no difference between the two groups, only that we don’t have good evidence that there is a difference.

Those in manual/routine occupations have blood pressure 3.22 mmHg higher than professional/managerial workers. By conventional criteria, the test statistic for this coefficient is t=3.77 and the p value <0.001, so we are more confident that we should reject the null hypothesis and conclude that there is a difference between the blood pressure of those in the manual/routine occupations compared to those in the professional/managerial groups.

We can use the linearHypothesis() function from car to carry out an F-test on multiple coefficients simultaneously. linearHypothesis() has a similar syntax to the glht() function from multcomp but does not need backticks to distinguish model terms.

linearHypothesis(mod\_2, c("sclass\_3Skilled Non-Manual = 0",  
 "sclass\_3Manual/Routine = 0"))

## Linear hypothesis test  
##   
## Hypothesis:  
## sclass\_3Skilled Non - Manual = 0  
## sclass\_3Manual/Routine = 0  
##   
## Model 1: restricted model  
## Model 2: sbp ~ sclass\_3  
##   
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 2634 1004133   
## 2 2632 998972 2 5161.5 6.7996 0.001134 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The p value is very low, so there is statistical evidence against the null hypothesis that all of the social class coefficients are 0. (This is not to say that **all** are not zero - only that at least one isn’t.)

## 2.3. Predicted values

**Q: What is the value of the estimated systolic blood pressure of each category of social class?**

coef(mod\_2)

## (Intercept) sclass\_3Skilled Non-Manual   
## 138.553362 1.418523   
## sclass\_3Manual/Routine   
## 3.224042

coef(mod\_2)[1] # Professional/managerial

## (Intercept)   
## 138.5534

coef(mod\_2)[1] + coef(mod\_2)[2] # Skilled non-manual

## (Intercept)   
## 139.9719

coef(mod\_2)[1] + coef(mod\_2)[3] # Manual/routine

## (Intercept)   
## 141.7774

Notice that these predictions give exactly the same results as the means we calculated above. Now we know that we cannot be as confident about the difference between the means of skilled non-manual group and professional/managerial group.

# 3. Multiple linear regression

In this section, we will run regression models with more than one predictor variables. We will use sbp as the dependent variable, and run a regression model using age, sex, social class, quintiles of household wealth (wealth5) and smoking (smok\_bin) as independent variables.

## 3.1. Examining variables

First of all we need to explore these variables and check the number of observations.

descr(elsa[c("age", "sbp")])

## Descriptive Statistics   
## elsa$Age in years   
## Label: Age in years   
## N: 3129   
##   
## age sbp  
## ----------------- --------- ---------  
## Mean 60.12 140.38  
## Std.Dev 9.60 19.65  
## Min 45.00 88.50  
## Q1 52.00 126.00  
## Median 59.00 138.50  
## Q3 67.00 151.50  
## Max 90.00 222.50  
## MAD 11.86 18.53  
## IQR 15.00 25.50  
## CV 0.16 0.14  
## Skewness 0.45 0.68  
## SE.Skewness 0.04 0.05  
## Kurtosis -0.71 0.66  
## N.Valid 3129.00 2692.00  
## Pct.Valid 100.00 86.03

freq(elsa[c("sex", "sclass\_3", "wealth5", "smok\_bin")])

## Frequencies   
## elsa$sex   
## Label: Sex   
## Type: Factor   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## ------------ ------ --------- -------------- --------- --------------  
## male 1368 43.72 43.72 43.72 43.72  
## female 1761 56.28 100.00 56.28 100.00  
## <NA> 0 0.00 100.00  
## Total 3129 100.00 100.00 100.00 100.00  
##   
## elsa$sclass\_3   
## Type: Factor   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## ------------------------ ------ --------- -------------- --------- --------------  
## Prof/Managerial 1078 35.23 35.23 34.45 34.45  
## Skilled Non-Manual 766 25.03 60.26 24.48 58.93  
## Manual/Routine 1216 39.74 100.00 38.86 97.79  
## <NA> 69 2.21 100.00  
## Total 3129 100.00 100.00 100.00 100.00  
##   
## elsa$wealth5   
## Label: quintiles of household wealth   
## Type: Numeric   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## ----------- ------ --------- -------------- --------- --------------  
## 1 541 17.58 17.58 17.29 17.29  
## 2 551 17.91 35.49 17.61 34.90  
## 3 576 18.72 54.21 18.41 53.31  
## 4 650 21.12 75.33 20.77 74.08  
## 5 759 24.67 100.00 24.26 98.34  
## <NA> 52 1.66 100.00  
## Total 3129 100.00 100.00 100.00 100.00  
##   
## elsa$smok\_bin   
## Label: Binary smoking status   
## Type: Factor   
##   
## Freq % Valid % Valid Cum. % Total % Total Cum.  
## -------------------- ------ --------- -------------- --------- --------------  
## never/ex occ 1375 43.94 43.94 43.94 43.94  
## ex/current reg 1754 56.06 100.00 56.06 100.00  
## <NA> 0 0.00 100.00  
## Total 3129 100.00 100.00 100.00 100.00

We need to run our regressions based on the sample with no missing data in all the variables listed above (This will be helpful for model comparision in the next section, as the sample used will be the same across models). We can use the function drop\_na() from tidyr (part of the tidyverse) to drop rows with missing values. Let’s save a new data frame with the smaller dataset we will be using in the regressions. Let’s also convert wealth5 to a factor variable so R knows to interpret this as a categorical, rather than continuous, variable.

elsa\_2 <- elsa %>%  
 select(age, sbp, sex, sclass\_3, wealth5, smok\_bin) %>%  
 drop\_na() %>%  
 mutate(wealth5 = factor(wealth5))

## 3.2. Multiple linear regression

Now run multiple linear regression. But let us start with 3 predictor variables: age, female, and social class. To add multiple independent variables, we simply add these to the model formula using the syntax y ~ x1 + x2 + ... + xn.

mod\_3 <- lm(sbp ~ sex + sclass\_3 + age, elsa\_2)  
summary(mod\_3)

##   
## Call:  
## lm(formula = sbp ~ sex + sclass\_3 + age, data = elsa\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.449 -12.984 -1.449 10.300 83.798   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 99.24499 2.33827 42.444 < 2e-16 \*\*\*  
## sexfemale -1.92465 0.76607 -2.512 0.01205 \*   
## sclass\_3Skilled Non-Manual 0.99266 0.99100 1.002 0.31659   
## sclass\_3Manual/Routine 2.51838 0.83640 3.011 0.00263 \*\*   
## age 0.67840 0.03772 17.987 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.37 on 2588 degrees of freedom  
## Multiple R-squared: 0.1177, Adjusted R-squared: 0.1164   
## F-statistic: 86.35 on 4 and 2588 DF, p-value: < 2.2e-16

confint(mod\_3)

## 2.5 % 97.5 %  
## (Intercept) 94.6599280 103.8300528  
## sexfemale -3.4268244 -0.4224734  
## sclass\_3Skilled Non-Manual -0.9505754 2.9359028  
## sclass\_3Manual/Routine 0.8783013 4.1584519  
## age 0.6044401 0.7523510

**Q: How do you interpret the output?**

When sex and social class are held constant, per 1-year increase in age, there is a 0.68 mmHg increase in systolic blood pressure. Compared to men; women have 1.92 mmHg lower blood pressure (when age and social class are held constant). Workers in manual/ routine jobs have 2.52 mmHg higher systolic blood pressure than professional/managerial workers (when age and sex are held constant). The difference in blood pressure for skilled non-manual and professional/managerial is small (0.99). Looking at its p-value and 95% CI, we can conclude that there is little clear statistical evidence against the null hypothesis of no difference between these two groups. The blood pressure for manual/routine workers is 2.52 mmHg higher than that of professional/managerial workers. Its p value is very low (0.003) and suggests low compatibility with the null hypothesis that there is no difference between manual/routine workers and professional/managerial workers.

**Q: What is the predicted systolic blood pressure for women aged 55 and works in a manual/ routine job?**

coef(mod\_3)

## (Intercept) sexfemale   
## 99.2449904 -1.9246489   
## sclass\_3Skilled Non-Manual sclass\_3Manual/Routine   
## 0.9926637 2.5183766   
## age   
## 0.6783955

coef(mod\_3)[1] + coef(mod\_3)[2] + coef(mod\_3)[4] + coef(mod\_3)[5]\*55

## (Intercept)   
## 137.1505

Alternatively, this could have been specified using glht().

glht(mod\_3,   
 "`(Intercept)` + 55\*`age` + `sexfemale` +   
 `sclass\_3Manual/Routine` = 0") %>%  
 confint()

##   
## Simultaneous Confidence Intervals  
##   
## Fit: lm(formula = sbp ~ sex + sclass\_3 + age, data = elsa\_2)  
##   
## Quantile = 1.9609  
## 95% family-wise confidence level  
##   
##   
## Linear Hypotheses:  
## Estimate  
## `(Intercept)` + 55 \* age + sexfemale + `sclass\_3Manual/Routine` == 0 137.1505  
## lwr   
## `(Intercept)` + 55 \* age + sexfemale + `sclass\_3Manual/Routine` == 0 135.7285  
## upr   
## `(Intercept)` + 55 \* age + sexfemale + `sclass\_3Manual/Routine` == 0 138.5724

Next, add wealth5 into the model.

mod\_4 <- lm(sbp ~ sex + sclass\_3 + age + wealth5, elsa\_2)  
summary(mod\_4)

##   
## Call:  
## lm(formula = sbp ~ sex + sclass\_3 + age + wealth5, data = elsa\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.149 -12.918 -1.323 10.604 84.700   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 101.45111 2.50771 40.456 < 2e-16 \*\*\*  
## sexfemale -2.04513 0.76630 -2.669 0.00766 \*\*   
## sclass\_3Skilled Non-Manual 0.64326 0.99932 0.644 0.51982   
## sclass\_3Manual/Routine 1.72410 0.88793 1.942 0.05228 .   
## age 0.66571 0.03818 17.437 < 2e-16 \*\*\*  
## wealth52 0.99934 1.25358 0.797 0.42541   
## wealth53 -1.15136 1.23294 -0.934 0.35048   
## wealth54 -1.74470 1.21693 -1.434 0.15178   
## wealth55 -2.21439 1.20486 -1.838 0.06619 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.35 on 2584 degrees of freedom  
## Multiple R-squared: 0.121, Adjusted R-squared: 0.1182   
## F-statistic: 44.45 on 8 and 2584 DF, p-value: < 2.2e-16

**Q: How do you interpret the coefficients of wealth?**

When age, sex, and social class are held constant, people in the highest quintile group (group 5) of household wealth had 2.21 mmHg lower systolic blood pressure than people in the lowest quintile group (group 1). Looking at its p value (0.066) and 95% CI, there is only some statistical evidence against the null hypothesis that there is no difference between group 5 and group 1. While people in the 2nd,3rd, and 4th household wealth group had a similar blood pressure level as people in the lowest quintile group (group 1), as their differences are less than 2 mmHg and their p values are large and 95% CIs are wide.

**Q: By comparing the models with and without wealth, what do their R-squared values tell us?**

In the current model, which means that 12% of the variability of systolic blood pressure is explained by our selected variables (age, female, social class and wealth). is slightly higher than the model without wealth (). However, we need caution in interpreting to justify the utility of a variable added into the model, as R2 increases even when the new variables have no real predictive capability.

The adjusted R-squared does this by adding a penalty to the R-squared for having extra variables in the models. We could alternatively look at the values of Root MSE (RMSE). Lower values of RMSE indicate better fit. Calculating these is simple. Within the model object (list) is a residuals vector. As the name suggests, RMSE is just need to take the square-root of the mean of the squared residuals.

sqrt(mean(mod\_3$residuals^2))

## [1] 18.34869

sqrt(mean(mod\_4$residuals^2))

## [1] 18.31522

Root MSE (18.315) is slightly lower than before (18.348).

**Q: Do you want to keep the wealth variable in your model?**

We can use the linearHypothesis() function again. Rather than write “wealth52 = 0” … “wealth55 = 0”, we can use paste0() to save time.

lincoms <- paste0("wealth5", 2:5, " = 0")  
lincoms

## [1] "wealth52 = 0" "wealth53 = 0" "wealth54 = 0" "wealth55 = 0"

linearHypothesis(mod\_4, lincoms)

## Linear hypothesis test  
##   
## Hypothesis:  
## wealth52 = 0  
## wealth53 = 0  
## wealth54 = 0  
## wealth55 = 0  
##   
## Model 1: restricted model  
## Model 2: sbp ~ sex + sclass\_3 + age + wealth5  
##   
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 2588 872997   
## 2 2584 869815 4 3182.5 2.3636 0.05098 .  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Results show that there is some statistical evidence (p=0.051) against the null hypothesis that all of the wealth coefficients are 0. In addition, R-squared and Root MSE show that the model with wealth is a slightly better model. Purely on statistical grounds, we may want to keep wealth in our model.

**Note that selecting variables purely on statistical grounds is not good practice - we also need to use theory to justify model choices. It is best that you are clear from the outset which variables you will use.**

Finally, add smok\_bin into the model. We can write the function across multiple lines to improve readability.

mod\_5 <- lm(sbp ~ sex + sclass\_3 + age +   
 wealth5 + smok\_bin, elsa\_2)  
summary(mod\_5)

##   
## Call:  
## lm(formula = sbp ~ sex + sclass\_3 + age + wealth5 + smok\_bin,   
## data = elsa\_2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.352 -12.975 -1.359 10.529 84.867   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 101.11058 2.55889 39.513 <2e-16 \*\*\*  
## sexfemale -1.94948 0.77954 -2.501 0.0125 \*   
## sclass\_3Skilled Non-Manual 0.63569 0.99949 0.636 0.5248   
## sclass\_3Manual/Routine 1.66505 0.89238 1.866 0.0622 .   
## age 0.66540 0.03819 17.426 <2e-16 \*\*\*  
## wealth52 1.02841 1.25446 0.820 0.4124   
## wealth53 -1.10483 1.23502 -0.895 0.3711   
## wealth54 -1.68832 1.21996 -1.384 0.1665   
## wealth55 -2.11775 1.21357 -1.745 0.0811 .   
## smok\_binex/current reg 0.50534 0.75370 0.670 0.5026   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.35 on 2583 degrees of freedom  
## Multiple R-squared: 0.1211, Adjusted R-squared: 0.118   
## F-statistic: 39.55 on 9 and 2583 DF, p-value: < 2.2e-16

confint(mod\_5)

## 2.5 % 97.5 %  
## (Intercept) 96.09288841 106.1282682  
## sexfemale -3.47806892 -0.4208859  
## sclass\_3Skilled Non-Manual -1.32418178 2.5955690  
## sclass\_3Manual/Routine -0.08481161 3.4149034  
## age 0.59052785 0.7402819  
## wealth52 -1.43144420 3.4882600  
## wealth53 -3.52656813 1.3169028  
## wealth54 -4.08051722 0.7038734  
## wealth55 -4.49742966 0.2619214  
## smok\_binex/current reg -0.97258071 1.9832517

sqrt(mean(mod\_5$residuals^2))

## [1] 18.31362

**Q: How to interpret the coefficients of smok\_bin?**

The difference in blood pressure between never/ex occasional smokers and ex-regular/current smokers is only 0.5 mmHg. Looking at its p value and 95% CI, we can conclude that there is little evidence of a difference between these two groups.

**Q: By comparing the models with and without smoking, what do their R-squared and Root MSE tell us?**

Adding smoking status to the model had little effect on R-squared and Root MSE.

**Q: Do you want to keep the smoking variable in your model?**

Purely on statistical grounds, we may want to exclude smoking from our final model, as it’s p-value is very large, and it did not improve the model fit. However, smoking has been identified from theory or prior evidence as an important risk factor for high blood pressure. So we may want to include it in our model even if its coefficient happens to have a large p-value in this particular analysis. Or we may want to use different measures of smoking status and to test its associate with sbp.

# 4. Optional exercise

**Q1: What is the mean BMI (bmi) of our sample? What is the median? Are they similar? How many observations do we have for BMI?**

**Q2: Is BMI normally distributed? How did you come to this decision?**

**Q3: Run a regression with continuous BMI as the outcome and age as the independent variable.**

* Examine the summary statistics first.
* What do the coefficient and confidence intervals mean?
* What does the tell you?
* Calculate the predicted sbp value for someone aged 60.

**Q4: Run a regression with continuous BMI as the outcome with ethnicity (ethni) as the independent variable.**

* Examine the summary statistics first.
* What do the coefficient and confidence intervals mean?
* What does the tell you?
* Calculate the predicted sbp value for someone whose ethnicity is non-white.

**Q5: Run a regression with BMI as the outcome (bmi) and self-reported general health (srh) as the independent variable. Recode srh into three categories (very good/good/ fair, bad or very bad).**

* Examine the summary statistics first.
* What do the coefficient and confidence intervals mean?
* What does the tell you?
* Calculate the predicted sbp value for someone who had fair, bad or very bad health.

**Q6: Run a multiple regression with BMI as the outcome and age, ethnicity, and self-reported general health as the independent variables**

* What do the coefficient and confidence intervals mean?
* What does the tell you?
* Calculate the predicted sbp value for someone aged 60, white, with good health.