Indicator Variables

So far, you have seen mostly regression models in which the predictor variables were quantitative. It is also possible to include categorical variables as predictors in regression models.

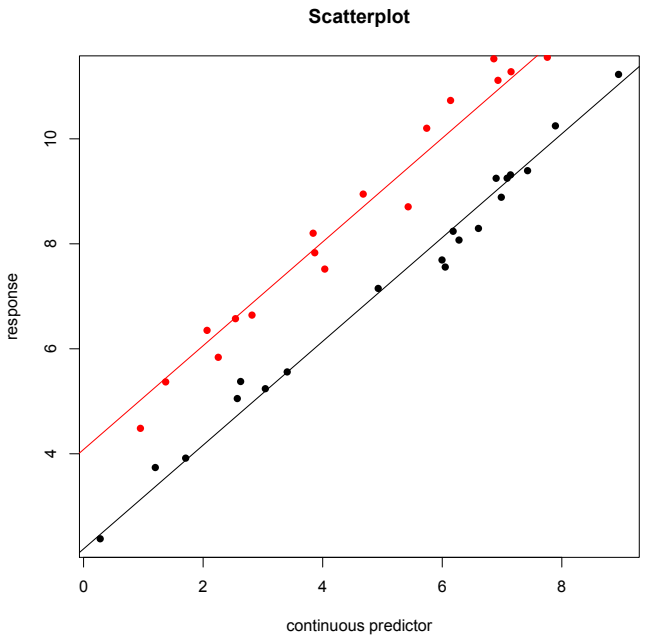
Example: Suppose that a quantitative response is to be modeled as a function of two predictors. x_1 is quantitative and x_2 is categorical and takes on the values A and B. We can code x_2 as follows:

$$x_2 = \begin{cases} 0 & \text{if observation is of type A} \\ 1 & \text{if observation is of type B} \end{cases}$$

and formulate the linear regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

Suppose the scatterplot below shows the observations on x_1 and y. The types are shown as different colors (A is black, B is red). Explain how you can see the estimated regression parameters $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ in the figure.



When the above linear model with a categorical predictor is fit, R will report a p-value for testing $H_0 : \beta_2 = 0$ vs. $H_a : \beta_2 \neq 0$. How do you interpret the outcome of this test?

The average height jumped by man has significant difference with women in same year
The increase of average height men and women can jump has significance difference with time goes by

H0: The average response y at the same x predictior level does not differ significantly for group A and B

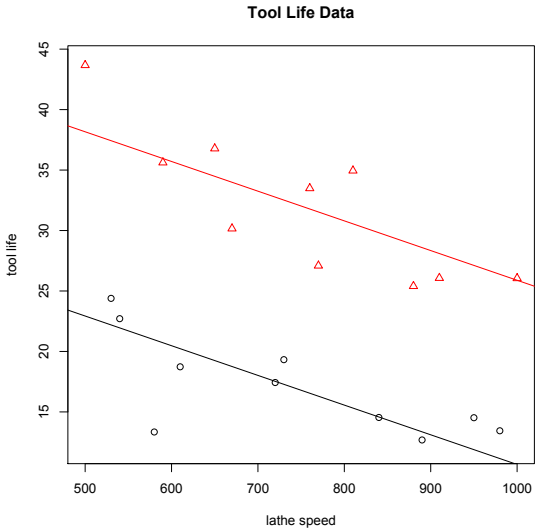
Explain why the two fitted regression lines for types A and B in the previous plot are parallel.

Because we give the same slope in both model

Note: There are other models - those with interaction terms - that allow for non-parallel lines.

Example: The Tool Life Data

A lathe is a machine which turns at a high rate and is used for machining metal pieces. Twenty observations on tool life (y in hours) and lathe speed (x_1 in rpm) have been collected for two different types of tools (A and B). The scatterplot of tool life against lathe speed clearly shows that two different regression lines are appropriate for the two types of tools (type A are shown as black dots and type B as red triangles).



To fit a multiple regression model in R, we first need to code an indicator variable for TOOL TYPE. Let

$$x_2 = \begin{cases} 0 & \text{if tool is of type A} \\ 1 & \text{if tool is of type B} \end{cases}$$

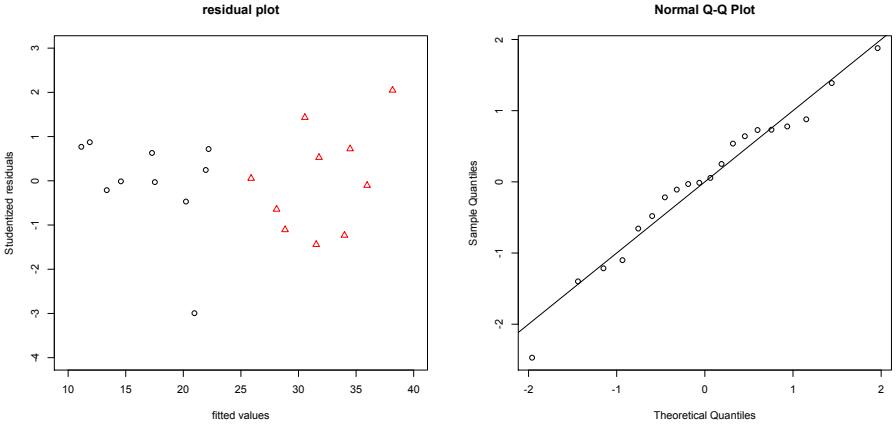
The summary for the linear model fit in R is shown below. Interpret the output.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.208726   3.738882   9.417 3.71e-08 ***
speed       -0.024557   0.004865  -5.048 9.92e-05 ***
dummy       15.235474   1.501220  10.149 1.25e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.352 on 17 degrees of freedom
Multiple R-squared: 0.8787, Adjusted R-squared: 0.8645
F-statistic:  61.6 on 2 and 17 DF,  p-value: 1.627e-08
```

t test on speed tell us:
the decrease in average tool life for both types of tools as lathe speed increases is significantly different from 0.

t test on dummy tell us:
the average tool life for type B is significantly larger than for type A if run at the same lathe speed

The residual plots for this model show that the residuals are approximately normally distributed (qq-plot). The residual plot shows that the residual variance for tools of type B (red triangles) seems to be larger than for tools of type A. That is a (mild) violation of the model assumptions.

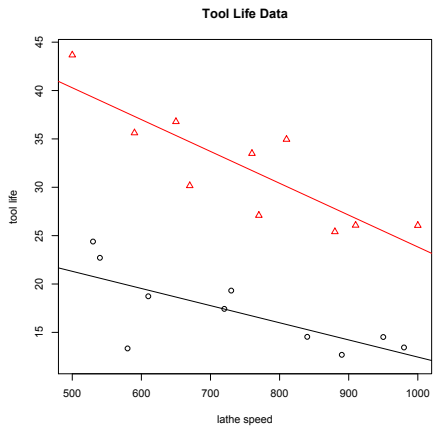


Could we have replaced the regression model we used with two simple linear regression models (one for each tool type)?

We could also allow the regression lines for the two different tool types in this example to have different slopes. This result could be achieved by fitting the linear regression model with interaction:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

Write down the two different regression lines this model uses for tools of type A & B:



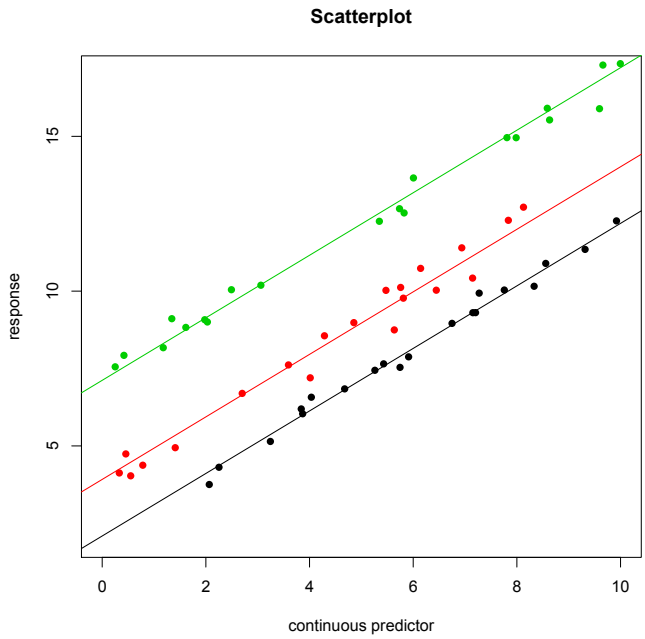
Categorical Predictors With More Than Two Levels

Suppose that a general categorical predictor has three levels: A,B, and C. We *should not* code the factor as

$$x_2 = \begin{cases} 0 & \text{if observation is type A} \\ 1 & \text{if observation is type B} \\ 2 & \text{if observation is type C} \end{cases}$$

Why?

If this is done in practice, it is called regression with allocated codes.



The linear regression model for one quantitative predictor and a categorical predictor with three levels (say A,B, and C) in which the three regression slopes are parallel can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Instead, when a categorical variable has more than two levels we will rely on INDICATOR VARIABLES which are also sometimes called DUMMY VARIABLES. The variables encode which two levels of the categorical factor are compared for each β-slope.

DEFINITION: A contrast is a linear combination of factor levels whose coefficients add up to zero. Each contrast corresponds to a testable hypothesis about the mean responses at the different factor levels.

Note: There are many different ways that indicator variables can be created to code for categorical predictors in regression. Each coding method creates a number of "dummy" variables that will be included in the linear regression model. The dummy variables are created to represent differences between factor levels. In general, if a categorical predictor in regression has a levels, we will need to create $a - 1$ dummy variables to include in the regression model.

Treatment Coding

One of the levels of the factor is chosen as a BASELINE. Both R and Pandas in Python choose the alphabetically first level as the baseline by default. Sometimes, it makes sense to use a different specific level (e.g., control group) as the baseline. The dummy variables then express the average changes in the response if the factor variable is changed from the baseline level to another level and all other variables are held fixed. The slopes corresponding to the indicator variables will be positive, if the average response of the factor level the indicator corresponds to is higher than for the baseline factor level. Otherwise the slope will be negative.

Creating the dummy variables by hand:

For treatment contrasts the dummy variables are indicator variables, that means that they take on only values of 0 and 1. We will create one less dummy variable than there are factor levels. There is always more than one way of encoding the dummy variables. Which way they are coded makes a difference for the interpretation of the slopes in the fitted regression model.

Example: Let's say you have a quantitative predictor x_1 and a categorical predictor with three levels (A,B, and C). Then we will need two dummy variables to encode this predictor. Let's call them x_2 and x_3. Pick one level of the predictor as a baseline (say, A). The baseline will correspond to values of 0 in both dummy predictors.

x_2	x_3	
0	0	if observation is of type A
1	0	if observation is of type B
0	1	if observation is of type C

If we give dummy x_2 a value of 1 for observations of type B and dummy x_3 a value of 1 for observations of type C (and the respective other dummy a value of zero), then the then the slope β_2 that corresponds to dummy x_2 can be interpreted as the average change in the response as you change the factor level from the baseline (A) to B and the slope β_3 for x_3 can be interpreted as the average change in the response if the factor level is changed from the baseline (A) to C.

Example: The R command for treatment contrast coding is `contr.treatment()`. This method is the default in R. Under the default option the (alphabetically) first factor level is selected as the baseline. You can designate other levels as the baseline with the argument `base`.

Understand the contrast coding and interpret the corresponding slopes for the following R-output. We are fitting the regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

where x_1 is a quantitative predictor, and x_2, x_3 are dummy variables that encode a categorical predictor with levels A, B, C.

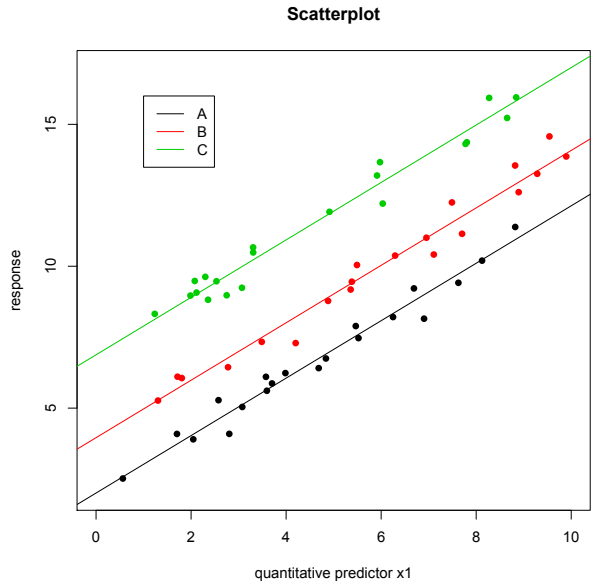
```
> contrasts(factor) <- contr.treatment(3, base = 1)
> contrasts(factor)
  2 3
A 0 0
B 1 0
C 0 1
> fit <- lm(y~x+factor)
> summary(fit)

Call:
lm(formula = y ~ x + factor)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91858 -0.31415 -0.00145  0.35370  0.95810

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.00797    0.15013   13.38   <2e-16 ***
x            1.01215    0.02379   42.55   <2e-16 ***
factor2      1.95080    0.14759   13.22   <2e-16 ***
factor3      4.87097    0.14437   33.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4565 on 56 degrees of freedom
Multiple R-squared: 0.9815, Adjusted R-squared: 0.9805
F-statistic: 988.1 on 3 and 56 DF,  p-value: < 2.2e-16
```

beta_1 !=0:

The quantitative predictor x has a significant influence on the response that is, the average y changes significantly as x increases.This change is assumed to be the same in all 3 groups for his model

beta_2 !=0:

The response means for groups B and A differs significantly at the same level of x



Models With More Than One Indicator Variable

Of course, in practice, in large regression applications there may often be more than one categorical variable (each with two or more levels) that are used as predictors in the same model. How are the regression slopes to be interpreted in this case?

Example: Suppose we have a continuous predictor x_1 and two categorical predictors with two factor levels, each. Let the levels of the first predictor be A_1, B_1 and the levels of the second predictor A_2, B_2. Since the predictors have two levels, each, we can use one indicator variable per predictor

$$x_2 = \begin{cases} 0 & \text{if factor 1 has level } A_1 \\ 1 & \text{if factor 1 has level } B_1 \end{cases}$$

$$x_3 = \begin{cases} 0 & \text{if factor 2 has level } A_2 \\ 1 & \text{if factor 2 has level } B_2 \end{cases}$$

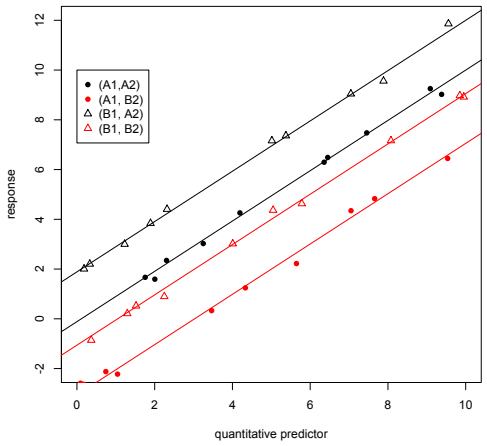
The simplest possible regression model for this problem then becomes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

Think about the regression lines representing the different subgroups in this model. What are the subgroups? What can you say about the slopes and intercepts of the individual regression lines?

In the above example, the four regression lines for the different combinations of factor levels $(A_1, A_2), (A_1, B_2), (B_1, A_2), (B_1, B_2)$ are all parallel. That means that they have the same slope (namely $\hat{\beta}_1$) and different intercepts.

A1A2: y=beta_0hat + beta_1hat x1
A1B2: y=beta_0hat+beta_3hat + beta_1hat x1
A2B1: y=beta_0hat+beta_2hat + beta_1hat x1
A2B2: y=beta_0hat+beta_1hat+beta_3hat + beta_1hat x1



Example: To allow the regression lines from the previous problem to have different slopes, we can add various interaction effects to the model:

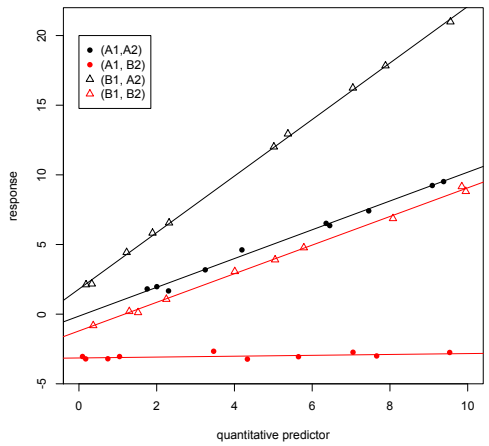
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \epsilon$$

What can you say about the slopes and intercepts of the four regression lines now?

A1A2: y=beta_0hat + beta_1hat x1
A1B2: y=beta_0hat+beta_3hat+(beta_1hat+beta_13)x1
A2B1: y=beta_0hat+beta_2hat+(beta_1hat+beta_12)x1
A2B2:y=A1B2: y=beta_0hat+beta_2+beta_3hat
+(beta_1hat+beta_13hat+beta_12hat)x1

Note, that the above model is still additive in the sense that a change from A_1 to B_1 changes the slopes and intercepts of the black regression lines. But the difference in intercepts of the two black lines and the two red lines is *the same*. And the difference in slopes between the two black and two red lines is also the same.

The rate increase faster as the x increases



Example: If we two more interaction terms to the model - namely the interaction between the two dummy variables and with the quantitative predictor, then we allow for the four lines to have four independent intercepts and slopes.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3 + \epsilon$$

Example: Interpret the meaning of the slope β_{123}.

The difference in angle between 2 black lines and two red lines respectely
also means the difference in the differences of the rates of change

Comparing Regression Models

Depending on the types of dummy variables used, we can test many different hypotheses about the effects that the changing levels of the factor variable(s) have on the response. But there are a couple of general questions that are of particular interest.

Parallel Lines: Suppose we have a factor with M levels and we have n_m observations on a quantitative predictor x_1 and a response y at each level. Then we can ask the question whether the individual slopes of the M factor levels are all the same, i.e.

$$y = \beta_{0m} + \beta_{1m}x_1 + \epsilon \quad m = 1, \ldots, M$$

and

$$H_0 : \beta_{11} = \beta_{12} = \cdots = \beta_{1M} \text{ vs. } H_a : \text{at least one slope is different from the others}$$

This question can be answered with an F-test in which the full model (M-separate regression equations) will be compared to the reduced model (one regression model with $M - 1$ dummy variables). The test statistic is, as usual

$$F_0 = \frac{\left((SS_{Res}(RM) - SS_{Res}(FM)\right)/(df_{RM} - df_{FM})}{SS_{Res}(FM)/df_{FM}} \sim F_{df_{RM}-df_{FM},df_{FM}}$$

where SS_{Res} of the full model may be found by adding the residual sums of squares of the M individual simple linear regression models. The degree of freedom of the full model is $df_{FM} = \sum(n_m - 2) = n - 2M$. The reduced model has $M - 1$ slopes for the dummy variables (plus 1 parameter each for the slope for the quantitative predictor and the intercept). Thus $df_{RM} = n - (M + 1)$.

Concurrent Lines: Do many regression lines all intersect in the same point? The full model remains the same (M individual simple linear regressions). But now we want to test whether the intercepts are the same (with possibly different slopes).

$$H_0 : \beta_{01} = \beta_{02} = \cdots = \beta_{0M} \text{ vs. } H_a : \text{at least one intercept is different from the others}$$

In this case the reduced model becomes

$$y = \beta_0 + \beta_1 x + \beta_2 x D_1 + \beta_3 x D_3 + \cdots + \beta_M x D_{M-1} + \epsilon$$

This reduced model also has $n - (M + 1)$ degrees of freedom.

Coincident Lines: In this most restrictive case we're asking for *both* the intercepts and slopes of all the individual regression models to coincide, i.e.

$$H_0 : \beta_{11} = \beta_{12} = \cdots = \beta_{1M} \textbf{ and } \beta_{01} = \beta_{02} = \cdots = \beta_{0M}$$

In this case, the reduced model is a simple linear regression scenario for all n data points at once.