Categorical predictors, interactions, polynomials

Emilija Perković

Dept. of Statistics
University of Washington

How to encode nominal predictors

We assume:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}$$

with $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$.
A column of the design matrix X represents values of predictor x_i.
How to encode nominal predictors (factors)? Examples:

- Smoking (yes/no)
- Type of tool (A,B,C)
- Dose of drug (I,II,III, IV)
- etc.

In R these are referred to as `factors` (see `as.factor()` in R).

Example - Cars (Acceleration)

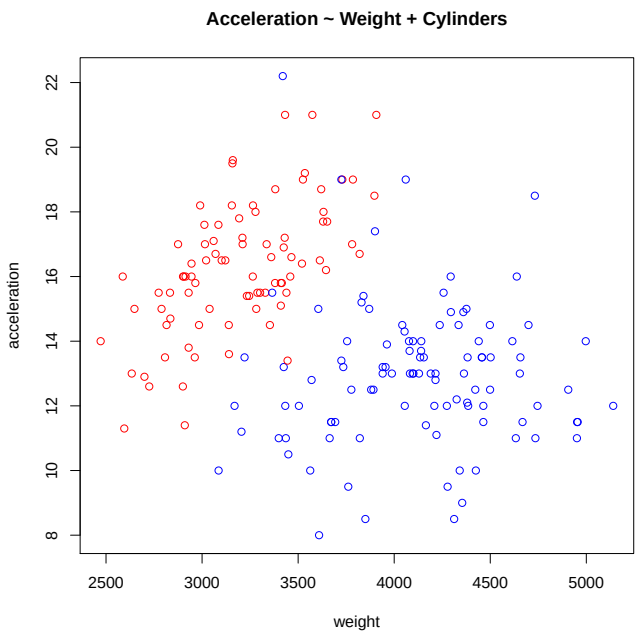
Let's start with a factor with 2 levels.
Dataset: `carsdata.RDS` on Canvas.
Some variables:

- weight - weight of a car,
- cylinders - number of cylinders (in this case 6, or 8),
- acceleration - how many sec. to reach 60 mph.

y - acceleration,
Predictors: x_1 - weight, x_2 - cylinders

Look at the scatter plot.

Example - Cars (Acceleration)



Example - Cars (Acceleration)

Let's start with a factor with 2 levels.
Dataset: `carsdata.RDS` on Canvas.
Some variables:

- weight - weight of a car,
- cylinders - number of cylinders (in this case 6, or 8),
- acceleration - how many sec. to reach 60 mph.

y - acceleration,
Predictors: x_1 - weight, x_2 - cylinders

How to encode the column corresponding to predictor x_2 in our design matrix X?
1) Use 6 and 8? 2) Use 2 columns? 3) Some other encoding?

Dummy encoding

We will consider "dummy" encoding of factor variables.
For predictor x_2 we let:

$$x_{i2} = \begin{cases} 1 & \text{if the number of cylinders is 8} \\ 0 & \text{otherwise} \end{cases}$$

And consider the model $\underline{y} = \beta_0 + \beta_1 \underline{x_1} + \beta_2 \underline{x_2} + \epsilon$.
How to interpret β_2 in the above linear regression?
What is the expectation of y if $x_2 = 0$?

$$\mathrm{E}[y|x_1 = x, x_2 = 0] = \beta_0 + \beta_1 x$$

What about if $x_2 = 1$?

$$\mathrm{E}[y|x_1 = x, x_2 = 1] = \beta_0 + \beta_1 x + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x$$

Both are straight lines, with the same slope and different intercepts!

In general, β_2 shows how much higher (lower) the mean response line is for the class coded 1 than the mean response line for the class coded 0, for any value of x_1. For us, $x_{i2} = 0$ if *cylinders* $= 6$, so *cylinders* $= 6$ is the *reference level*.

Why not other encodings?

Why didn't we choose to use 6 and 8 instead of 0 and 1?
Using this encoding:

$$\mathrm{E}[y|x_1 = x, x_2 = 6] = \beta_0 + \beta_1 x_1 + 6\beta_2 = \beta_0^1 + \beta_1 x$$

$$\begin{aligned}\mathrm{E}[y|x_1 = x, x_2 = 8] &= \beta_0 + \beta_1 x + 8\beta_2 \\ &= (\beta_0^1 + 2\beta_2) + \beta_1 x = (\beta_0^1 + \beta_2^1) + \beta_1 x,\end{aligned}$$

where $\beta_0^1 = \beta_0 + 6\beta_2$ and $\beta_2^1 = 2\beta_2$.

The dummy encoding is essentially equivalent to using the encoding with 6 and 8.

The reason we use the dummy encoding is to have a more general framework when working with other categorical variables that do not have a numerical representation, e.g., Types of drugs: A, B, etc.

Why not other encodings?

Why didn't we choose to use two columns in the design matrix? For example:

$$(\underline{cylinders6}, \underline{cylinders8}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix}$$

In this case, the sum of these two columns of the design matrix would be $(1, \ldots, 1)'$ (same as the intercept), meaning that the two columns of the design matrix would be linearly dependent with the intercept column.

Remember issue of multicolinearity, from MLR I lecture. In this case, we cannot estimate $\underline{\hat{\beta}}$.
We still use OLS to estimate $\underline{\beta}$, that is still $\underline{\hat{\beta}} = (X'X)^{-1}X'y$.

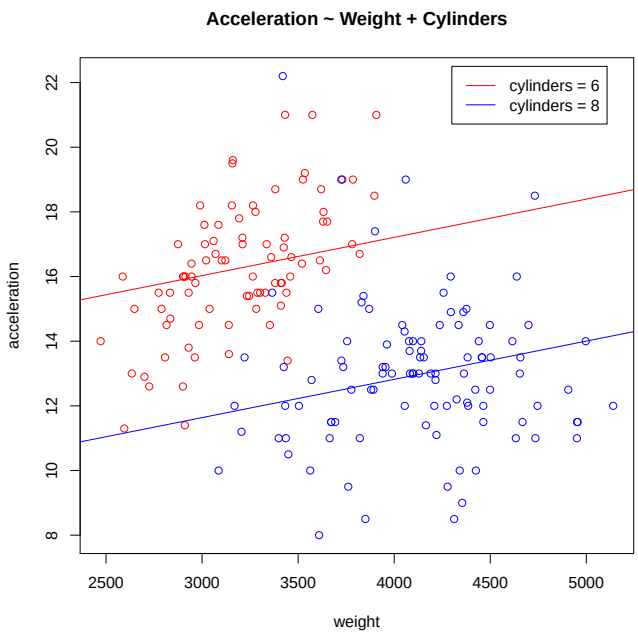
Example - Acceleration

```
lm(formula = acceleration ~ weight + cylinders,
   data = cars.data1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.4782362  1.2461692  10.013  < 2e-16
weight       0.0011834  0.0003831   3.089  0.00232
cylinders1  -4.3923548  0.4663726  -9.418  < 2e-16
---
Residual standard error: 2.088 on 184 degrees of freedom
Multiple R-squared:  0.4058,Adjusted R-squared:  0.3993
F-statistic: 62.82 on 2 and 184 DF,  p-value: < 2.2e-16
```

The fitted acceleration for car with *weight* $= 3000$ and 6 cylinders is ≈ 16.029. What would be the fitted acceleration for a car with the same weight and 8 cylinders? $\approx 16.029 - 4.39 = 11.639$.

Example - Acceleration



Interpretation and inference

Note that from previous slide:
$\beta_2 = \mathrm{E}[y|x_1 = x, x_2 = 1] - \mathrm{E}[y|x_1 = x, x_2 = 0]$.

Interpretation of β_2:

- The average change in y for two cars that are identical in weight when changing the number of cylinders from 6 to 8.

The **hypothesis tests** and **confidence intervals** for β_2 can be performed and calculated in the same way as discussed in the previous lectures.

Why not split the data according to the number of cylinders and perform two regressions?

- If you have enough samples for both cylinder types, than can be a reasonable thing to do.
- However, splitting the data will affect the estimate of the error variance. Hence, your tests will be less powerful.

Categorical predictors with k levels

What if a categorical variable $\tilde{x}_l$ has $\{1, 2, \ldots, k\}$ levels, $k > 2$? How do we include this predictor in our linear regression model?
Use $k-1$ dummy variables $x_1, \ldots, x_{k-1}$ to encode it. For example for $j \in \{1, \ldots, k-1\}$, dummy variable $x_j \in \{0, 1\}$:

$$x_{ij} = \begin{cases} 1 & \text{if } \tilde{x}_{il} = j + 1 \\ 0 & \text{otherwise} \end{cases}$$

Let's go back to a modified version of our example:
Dataset: `carsdata.RDS` on Canvas.
Some variables:

- weight - weight of a car,
- cylinders - number of cylinders (in this case **4, 6, or 8**),
- acceleration - how many sec. to reach 60 mph.

y - acceleration, x_1 - weight, $\tilde{x}_2$ - cylinders
How to encode cylinders?

Categorical predictor with 3 levels

Now, our predictor $\tilde{x}_2$ has 3 levels. We encode this with 2 dummy predictors x_2 and x_3:

$$x_{i2} = \begin{cases} 1 & \text{if the car has 6 cylinders} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{if the car has 8 cylinders} \\ 0 & \text{otherwise} \end{cases}$$

Our reference level is now 4 cylinders.
(We could have also chosen a different reference level.)
We fit the model: $\underline{y} = \beta_0 + \beta_1 \underline{x_1} + \beta_2 \underline{x_2} + \beta_3 \underline{x_3} + \epsilon$.
We now have:

$$\begin{aligned} \mathrm{E}[y|x_1 = x, x_2 = 0, x_3 = 0] &= \beta_0 + \beta_1 x \\ \mathrm{E}[y|x_1 = x, x_2 = 1, x_3 = 0] &= \beta_0 + \beta_1 x + \beta_2 = (\beta_0 + \beta_2) + \beta_1 x \\ \mathrm{E}[y|x_1 = x, x_2 = 0, x_3 = 1] &= \beta_0 + \beta_1 x + \beta_3 = (\beta_0 + \beta_3) + \beta_1 x \end{aligned}$$

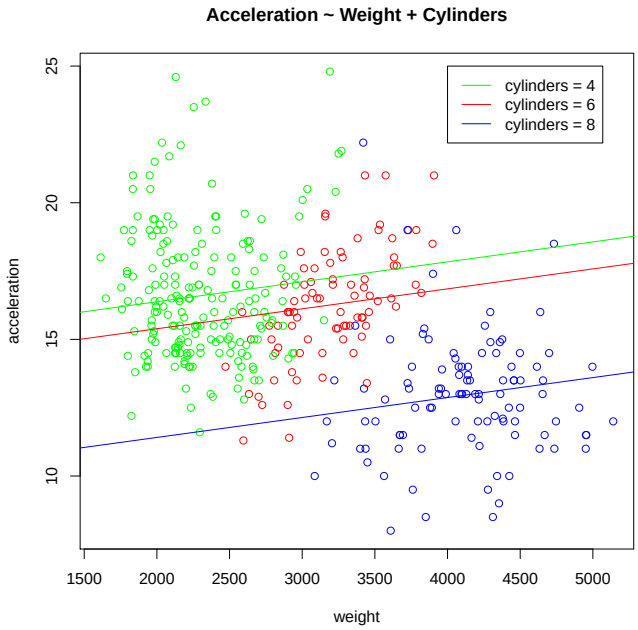
Example: Acceleration continued

As long as cylinders is encoded as a factor in R (hint: `as.factor()`), you can fit the linear model with `lm(.)` as before. (What happens if it is not encoded as a factor?) Again $\hat{\underline{\beta}} = (X'X)^{-1}X'y$.

```
lm(formula = acceleration ~ weight + cylinders,
   data = cars.data)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.9164014  0.7244317  20.590  < 2e-16
weight       0.0007301  0.0003063   2.383   0.0176
cylinders6  -0.9881999  0.3997077  -2.472   0.0139
cylinders8  -4.9650488  0.6168546  -8.049 1.04e-14
---
```

Fitted value of fuel for a car with weight 3000 and 6 cylinders is: ≈ 16.118.
What is the fitted value of fuel for a car with weight 3000 and 8 cylinders?
$\approx 16.118 + 0.988 - 4.965 = 12.141$. Check in R.

Example: Acceleration continued



Interpretation of coefficients

How do we interpret β_2 and β_3 in this fit?

- β_2 - The average change in y for two cars that are identical in weight when changing the number of cylinders from 4 to 6.
- β_3 - The average change in y for two cars that are identical in weight when changing the number of cylinders from 4 to 8.

The estimated coefficients $\hat{\beta}_2$, $\hat{\beta}_3$ are calculated in reference to 4 cylinders. The same is true for p-values in the hypothesis tests $H_0^1 : \beta_2 = 0$ and H_0^2: $\beta_3 = 0$.
What if we had used 6 cylinders as a reference level?
For example: $x_{i2} = 1$ if the car has 4 cylinders, and 0 otherwise, and $x_{i3} = 1$ if a car has 8 cylinders and 0 otherwise (in R see `relevel()`).

Example - Acceleration, different reference level

```
lm(formula = acceleration ~ weight + cylinders,
    data = cars.data2)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.9282014  1.0100913  13.789   <2e-16
weight       0.0007301  0.0003063   2.383   0.0176
cylinders4   0.9881999  0.3997077   2.472   0.0139
cylinders8  -3.9768489  0.4343643  -9.156   <2e-16
```

The estimated coefficient for *cylinders*8 has changed! It used to be:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.9164014  0.7244317  20.590  < 2e-16
weight       0.0007301  0.0003063   2.383   0.0176
cylinders6  -0.9881999  0.3997077  -2.472   0.0139
cylinders8  -4.9650488  0.6168546  -8.049 1.04e-14
```


Practical implications

This means that our summary output is **tied to our choice of reference level**!

If one or more of the coefficients corresponding to dummy predictors is not significant using one reference level, they may be significant when using another reference level!

How to perform hypothesis tests? Remove a few levels?

For categorical predictors it is generally recommended to either include all levels of the categorical predictor or do not include the categorical predictor in the model.
The appropriate test to perform is the anova test comparing a full and a reduced model.
Null hypothesis? Test statistic? Distribution of the test statistic under the null?

Recall: comparing models.

We compare the fits of model

$$\underline{y} = \beta_0 + \beta_1 \underline{x}_1 + \beta_2 \underline{x}_2 + \beta_3 \underline{x}_3 + \underline{\epsilon} \quad (1)$$

and model

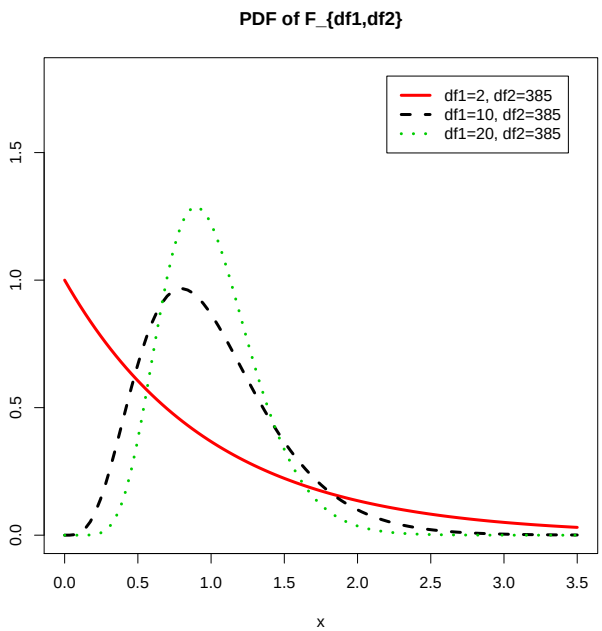
$$\underline{y} = \beta_0 + \beta_1 \underline{x}_1 + \underline{\epsilon} \quad (2)$$

to test $H_0 : (\beta_2, \beta_3)' = (0, 0)$ against $H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$. In our case: $n = 391$.
We use the F-statistic

$$F = \frac{(RSS_1 - RSS_3)(391 - 3 - 1)}{(3-1)RSS_3},$$

where RSS_1 is the residual sum of squares after fitting model (1) and RSS_3 is the the residual sum of squares after fitting model (2). If H_0 is true, $F \sim F_{2,387}$.

Recall: F-distribution



F-test

$P(F < 0) = 0$, if $F \sim F_{2,387}$. Use one sided p-values and confidence intervals!
If f is the observed value of

$$F = \frac{(RSS_1 - RSS_3)387}{2RSS_3}$$

.

Then the p-value for testing $H_0 : (\beta_2, \beta_3)' = (0, 0)$ against $H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$ is calculated as

$$P(F \geq f), \text{ where } F \sim F_{2,387}$$

In R, pf(f,df1=2,df2=387).

Example - Acceleration, anova

```
Analysis of Variance Table

Model 1: acceleration ~ weight
Model 2: acceleration ~ weight + cylinders
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    389 2406.7
2    387 1967.0  2    439.63 43.247 < 2.2e-16

> ## doing the F-test (anova test) by hand
> f <- (2406.7-1967)/1967*(387/2)
> 1-pf(f,2,387)
[1] 0
```

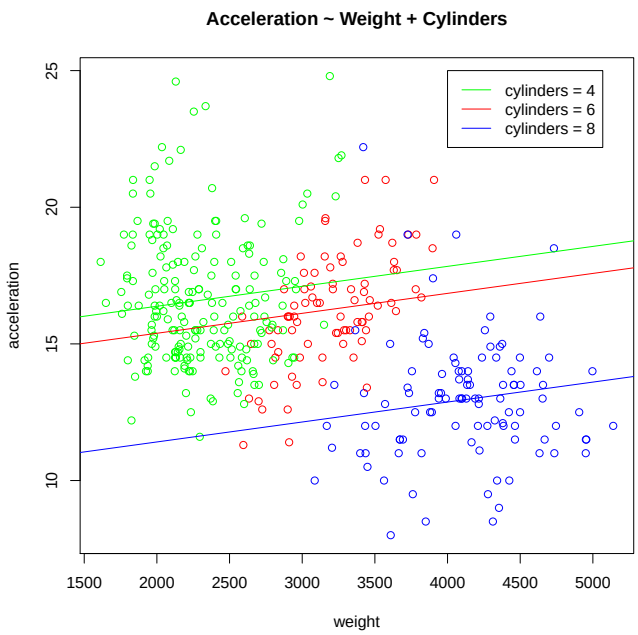
Decision: Reject $H_0 : \beta_2 = \beta_3 = 0$. We prefer the model that includes cylinders as a predictor.
The result (p-value) of the test is the same regardless of the reference level encoding.

Categorical predictors with *k* levels

A word of caution for predictors with many levels:

- In order to to fit a model that includes a categorical predictor with *k* levels, each of the *k* levels should be represented in the data sample.
- If this is not the case, perhaps make a new predictor by combining some levels of the original predictor.
- Depends on the data and the research question of interest.

Why not fit different slopes?



Interactions

On the previous slide we saw the fitted lines of the following model:
$\underline{y} = \beta_0 + \beta_1 \underline{x_1} + \beta_2 \underline{x_2} + \beta_3 \underline{x_3} + \epsilon$.

If we want to model different slopes for the different levels of cylinders we should include the interaction of weight and cylinder.

Our model is then:

$$\underline{y} = \beta_0 + \beta_1 \underline{x_1} + \beta_2 \underline{x_2} + \beta_3 \underline{x_3} + \beta_4 \underline{x_1 x_2} + \beta_5 \underline{x_1 x_3} + \epsilon$$

So:

$$\begin{aligned}
&E[y|x_1 = x, x_2 = 0, x_3 = 0] = \beta_0 + \beta_1 x \\
&E[y|x_1 = x, x_2 = 1, x_3 = 0] = \beta_0 + \beta_1 x + \beta_2 + \beta_4 x_1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x \\
&E[y|x_1 = x, x_2 = 0, x_3 = 1] = \beta_0 + \beta_1 x + \beta_3 + \beta_5 x_1 = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x
\end{aligned}$$

Again we use OLS to fit $\underline{\hat{\beta}} = (X'X)^{-1}X'y$. Fitting in R: see R script.

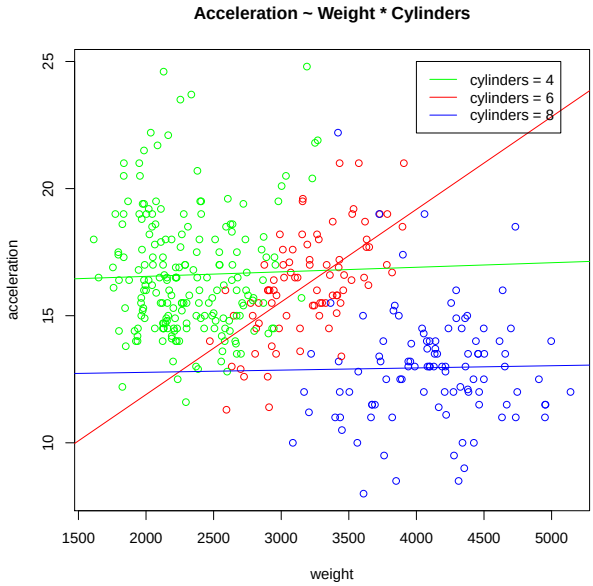
Example: Acceleration, interactions

```
lm(formula = acceleration ~ weight * cylinders,
  data = cars.data)
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.619e+01  1.041e+00  15.549  < 2e-16
weight             1.792e-04  4.460e-04   0.402    0.688
cylinders6        -1.158e+01  2.564e+00  -4.515 8.42e-06
cylinders8        -3.588e+00  2.267e+00  -1.583    0.114
weight:cylinders6  3.464e-03  8.543e-04   4.054 6.09e-05
weight:cylinders8 -9.278e-05  6.601e-04  -0.141    0.888
```

Alternatively:

```
lm(formula = acceleration ~ weight + cylinders
 + weight:cylinders,
  data = cars.data)
```

Example: Acceleration, interactions



Consider which interaction terms were significant in the summary! What if 6 was the reference level?

Example: Acceleration, interactions

```
lm(formula = acceleration ~ weight * cylinders,
  data = cars.data2)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       4.6125365  2.3427578   1.969   0.0497
weight            0.0036428  0.0007286   4.999 8.75e-07
cylinders4       11.5752139  2.5636551   4.515 8.42e-06
cylinders8        7.9870202  3.0895105   2.585   0.0101
weight:cylinders4 -0.0034636  0.0008543  -4.054 6.09e-05
weight:cylinders8 -0.0035564  0.0008762  -4.059 5.97e-05
---
Residual standard error: 2.206 on 385 degrees of freedom
Multiple R-squared:  0.3666,Adjusted R-squared:  0.3583
F-statistic: 44.56 on 5 and 385 DF,  p-value: < 2.2e-16
```

Now all coefficients are significant!

Interpretation of coefficients

In the model:

$$\underline{y} = \beta_0 + \beta_1 \underline{x_1} + \beta_2 \underline{x_2} + \beta_3 \underline{x_3} + \beta_4 \underline{x_1 x_2} + \beta_5 \underline{x_1 x_3} + \epsilon$$

it is no longer correct to interpret:

- $\beta_1 \stackrel{!}{=} \mathrm{E}[y|x_1 = x+1, x_2 = 1, x_3 = 0] - \mathrm{E}[y|x_1 = x, x_2 = 1, x_3 = 0]$

Since now:

$$\mathrm{E}[y|x_1 = x+1, x_2 = 1, x_3 = 0] - \mathrm{E}[y|x_1 = x, x_2 = 1, x_3 = 0] = \beta_1 + \beta_4.$$

We instead say:

- β_1 is the slope with regard to x_1 when $x_2 = 0$ and $x_3 = 0$, or
- β_1 is the "main effect" of x_1.
- while β_4 is the "interaction effect" of $x_1 x_2$.

Estimation and testing

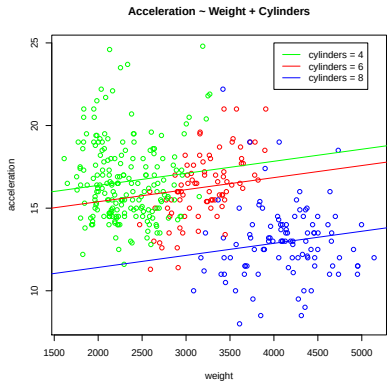
Estimation of $\underline{\beta}$ still follows from $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$.
Hypothesis testing for the interaction follows the same principle as for the main effect of a categorical variable.
In general, we will either be including all levels of the interaction or none:

- use the `anova()` test to compare the model with the interaction term and the model without the interaction term.
- Null hypothesis: $H_0 : \beta_4 = \beta_5 = 0$, ($H_1 : \beta_4 \neq 0$ or $\beta_5 \neq 0$),
- In this case: $q = 3$, $p = 5$, $n = 391$, and the test statistic $\frac{RSS_3 - RSS_5(391-5-1)}{RSS_5(5-3)}$ follows $F_{2,385}$ distribution under H_0.

```
Analysis of Variance Table

Model 1: acceleration ~ weight + cylinders
Model 2: acceleration ~ weight * cylinders
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    387 1967.0
2    385 1873.4  2    93.689 9.6271 8.321e-05
```

When to include an interaction



- Use scatter plots as indication. Add colors to differentiate groups.
- Consider the sample size and the number of parameters being estimated.
- Generally, having a hierarchical model is recommended - only include the interaction x_1x_2 if the main effects x_1 and x_2 are in the model. Otherwise, strange dependence of y on, for example x_1 (only sensitive to x_1 for $x_2 = 0$).

Continuous and higher order interactions

We can also consider interactions between continuous predictors, between two categorical predictors, or higher order interactions.

Let's consider some other predictors in our `cars.data` data set.

- weight - weight of a car,
- cylinders - number of cylinders (in this case 4, 6, or 8),
- acceleration - how many sec. to reach 60 mph,
- horsepower - horsepower of a car,
- mpg - miles per gallon consumption of a car.

y - acceleration, x_1 - weight, x_2, x_3 - cylinders (as before),
x_4 - horsepower, x_5 - mpg.

Example - Acceleration

y - acceleration, x_1 - weight, x_2, x_3 - cylinders (as before),
x_4 - horsepower, x_5 - mpg.
We can choose to fit a multiple linear regression with

- all main effects, and
- all two way interactions (interaction of two main effects), and
- the three-way interaction of weight, horsepower and mpg, that is,

$$\begin{aligned}\underline{y} = \beta_0 + \beta_1\underline{x}_1 + \beta_2\underline{x}_2 + \beta_3\underline{x}_3 + \beta_4\underline{x}_4 + \beta_5\underline{x}_5 \\ + \beta_6\underline{x}_1\underline{x}_2 + \beta_7\underline{x}_1\underline{x}_3 + \beta_8\underline{x}_1\underline{x}_4 \\ + \cdots + \beta_{14}\underline{x}_4\underline{x}_5 + \beta_{15}\underline{x}_1\underline{x}_4\underline{x}_5 + \underline{\epsilon},\end{aligned}$$

again, $\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$.
See R script for fit and output.

Polynomial predictors

We can fit even more complicated models by including, for example, $\underline{x}_1^2$ as a column to our design matrix.
Suppose y is not linearly related to x_1, but rather polynomially related, with polynomial of degree d (example: $d = 3$).

We may consider fitting:

$$\begin{aligned}\underline{y} = \beta_0 + \beta_1\underline{x}_1 + \beta_2\underline{x}_2 + \beta_3\underline{x}_3 + \beta_4\underline{x}_4 + \beta_5\underline{x}_5 \\ + \beta_6\underline{x}_1^2 + \beta_7\underline{x}_1^3 + \beta_7\underline{x}_1\underline{x}_2 + \beta_8\underline{x}_1\underline{x}_3 + \beta_9\underline{x}_1\underline{x}_4 \\ + \cdots + \beta_{16}\underline{x}_4\underline{x}_5 + \beta_{17}\underline{x}_1\underline{x}_4\underline{x}_5 + \underline{\epsilon},\end{aligned}$$

$\underline{\epsilon} \sim \mathcal{N}(\underline{0}, \sigma^2 I_n)$.

- Note that this does not cause issues with the estimation of $\underline{\beta}$ since $\underline{x}_1^2$ and $\underline{x}_1^3$ are not linear functions of x_1.
  See R script for how to fit data.

Polynomial predictors

Even though including polynomials of degree d does not cause problems with the estimation of $\hat{\underline{\beta}}$, $\underline{x}_1$ and $\underline{x}_1^2$ can be highly correlated.
See R script.

The same holds for interactions of continuous predictors, $\underline{x}_1$, can we highly correlated with $\underline{x}_1\underline{x}_4$.
One way to remedy this is to **center** your continuous predictors before including polynomials of a higher degree or interactions. Use $\underline{x}_1 - \overline{\underline{x}_1}$. See R script.
How to interpret the output?

Another way is to use the R function `poly()` which uses orthogonal basis functions to construct orthogonal polynomials. See R script, `help(poly)` and Cosma Shalizi's lecture notes (linked on Canvas). Interpretation?

Things to consider

- How many parameters are you estimating and how many samples do you have?
- What domain does the data come from and is there any evidence of some polynomial or interaction relationships.
- Interpretation. How do you interpret each coefficient in your model?
- The p-value estimates from the multiple linear regression fit summary are overly optimistic.
- Issues of overfitting and how to deal with them. To be covered later in the course.

General guideline: Parsimony.
If two models are explaining the data equally well, then opt for the one containing fewer predictor variables.
What does it mean for the model to explain the data well?
One idea: Use $R^2 = 1 - \frac{RSS}{SYY} = 1 - \frac{RSS}{\sum(y_i - \bar{y})^2}$ (coefficient of determination).
Or $\hat{\sigma}$, or MSE. See R script.
Other measures (to be covered later in the course).