INTERACTION EFFECTS

- What are interaction effects?

  [a] Two or more exogenous variables ***do not act independently*** on the dependent variable. That is, a ***direct interpretation*** of the slope parameter becomes invalid: $\Delta\hat{Y} \neq b_k \cdot \Delta X_k$.

  [b] An interaction is ***more than*** the simple ***sum*** of impacts from a set of independent variables.

  [c] The ***effect of one variable*** changes systematically with the ***level of another variable***.

  [d] Consequently, the slope coefficient associated with one variable becomes dependent on the ***level*** of another variable.

  Consequence: Regression parameters can no longer be interpreted as single partial effects, because their impact dependents on the level of other variables.

  Conditional effect plots, though, can be used to visualize the effect (see **HAM** p 161 and the `effects` library).

- Discuss example in **HAM** pp 84-85 for the slope coefficient of ΔX :

  Let $\hat{Y}_i = b_0 + b_1 \cdot X_i + b_2 \cdot W_i + b_3 \cdot (X_i \cdot W_i)$

  Then any change ΔX leads to a change in $\Delta\hat{Y}$ that is

$$\Delta\hat{Y} = (b_1 + b_3 \cdot W_i) \cdot \Delta X$$

Thus the change in $\Delta\hat{Y}$ depends on the given level of W_i

If W is:	the slope coefficient of X becomes
$W_i = 0$	$b_1 + b_3 \cdot 0 = b_1$
$W_i = 1$	$b_1 + b_3 \cdot 1 = b_1 + b_3$
$W_i = 2$	$b_1 + 2 \cdot b_3$

If b_1 and b_3 have ***opposite signs*** the direction of the relationship between X and Y may change depending on the given level of W_i.

- ***Products*** of the independent variables lead to interaction effects. See script **`attend.r`** and **`model.matrix( )`** function.
- Interaction effects increase the ***risk of multicollinarity*** because the product of two independent variables shares common information from both parent variables.

- Interaction effects of a variable with itself, e.g., polynomial functions, also allows to model ***non-linear relationships*** between the dependent and independent variables.
  See, for instance, a quadratic function with $\hat{y}_i = b_0 + b_1 \cdot x_i + b_2 \cdot x_i^2$ and the script **`hPrice2.r`**.

- In the ***Cobb-Douglas*** economic production function all independent variables measure implicitly interaction:

$$y_i = b_0 \cdot x_{i1}^{b_1} \cdot x_{i2}^{b_2} \cdot e_i$$

  The highly non-linear model can be transformed into a linear form by the log-transformation:

$$\Rightarrow \log(y_i) = \log(b_0) + b_1 \cdot \log(x_{i1}) + b_2 \cdot \log(x_{i2}) + \log(e_1)$$

  under the assumptions that ε_i is i.i.d. ***log-normal*** distributed, then $\log(\varepsilon_i)$ become ***normal*** distributed. The positivity condition $y_i > 0, b_0 > 0, x_{i1} > 0$ and $x_{i2} > 0$ must hold in order to apply the logarithmic function.

INDICATOR VARIABLES IN REGRESSION ANALYSIS

1. Dummy variables allow the inclusion of a ***categorical independent variable***.

   $\Rightarrow$ independent variables do not need to be measured on a metric level.

   Note: For categorical independent variables transformations are meaningless.

   In the *regression analysis* and the *analysis of variance* terminology a categorical variable is called a ***factor*** and operationalized by a set of ***indicator*** variables.

2. A categorical variable with ***more than two levels*** is ***coded*** into several ***indicator*** variables.

3. Example for "*one-hot encoding*" with three ($J = 3$) ***mutually exclusive*** categorical levels in a factor:

$$D_{i1} = \begin{cases} 1 & \text{if observation } i \text{ belongs to category 1} \\ 0 & \text{if observation } i \text{ does not belong to category 1} \end{cases}$$

$$D_{i2} = \begin{cases} 1 & \text{if observation } i \text{ belongs to category 2} \\ 0 & \text{if observation } i \text{ does not belong to category 2} \end{cases}$$

$$D_{i3} = \begin{cases} 1 & \text{if observation } i \text{ neither belongs to category 1 or 2} \\ 0 & \text{if observation } i \text{ belongs to category 1or category 2} \end{cases}$$

This would lead to a regression system, which becomes depending on an observation's classification:

[a] for belonging to category 1: $\hat{y}_i = b_0 + b_1 \cdot 1 + b_2 \cdot 0 + b_3 \cdot 0$

[b] for belonging to category 2: $\hat{y}_i = b_0 + b_1 \cdot 0 + b_2 \cdot 1 + b_3 \cdot 0$

[c] for belonging to category 3: $\hat{y}_i = b_0 + b_1 \cdot 0 + b_2 \cdot 0 + b_3 \cdot 1$

4. ***Only conceptionally*** the *one-hot encoding* can be used in regression analysis. There are as many dummy variables for a categorical variable as there are categorical levels. In practice, one category is ***redundant*** and needs to be dropped from the analysis. This leads to the "*dummy encoding*".

   Note: The ***full set of J dummy variables is multicollinear*** with the intercept: $1 = D_{i1} + D_{i2} + D_{i3}$.

   - Thus ***either*** the intercept ***or*** a single dummy variable, associate with one category (usually the last one), needs to be dropped from the model specification.
   - After dropping the dummy variable D_{i3} associated with last categorywe get therefore:

[a] for category 1: $\hat{y}_i = b_0 + b_1 \cdot 1 + b_2 \cdot 0$

[b] for category 2: $\hat{y}_i = b_0 + b_1 \cdot 0 + b_2 \cdot 1$

[c] for category 3: $\hat{y}_i = b_0 + b_1 \cdot 0 + b_2 \cdot 0$

5. Important: Since we estimate $J - 1$ parameters for a particular factor, these parameters need to be ***tested simultaneously*** with a ***partial* F-*test*** in order to decide whether the ***entire factor is relevant***.

6. Retirement in the water consumption example is a dummy variable

   - For non-retired households $(X_4 = 0)$the regression equation becomes:

     $\hat{Y}_i = 242 + 21 \cdot X_{i1} + 0.49 \cdot X_{i2} - 42 \cdot X_{i3} + 189 \cdot 0 + 248 \cdot X_{i5} + 96 \cdot X_{i6}$.

   - Whereas, for retired households $(X_4 = 1)$ the equation becomes:

     $\hat{Y}_i = 242 + 21 \cdot X_{i1} + 0.49 \cdot X_{i2} - 42 \cdot X_{i3} + 189 \cdot 1 + 248 \cdot X_{i5} + 96 \cdot X_{i6}$.

     Thus for retired households the intercept becomes: $431 = 242 + 189$.

INTERACTION BETWEEN INDICATOR AND METRIC VARIABLES IN REGRESSION ANALYSIS

7. Indicator variables allow us to model ***different regression regimes*** simultaneously and to cope with ***model heterogeneity***.

8. Each regime has its own ***intercept***, ***slope*** or even ***both***.

9. However, all regression regimes must still satisfy the model assumptions of OLS. In particular, that the ***disturbance*** for each regime of observation have same distributional properties, such as, identical variances.

EXAMPLE: HAMILTON'S WELLS IN DEPENDENCE OF DISTANCE AND SHALLOWNESS (PP 86-92):

- This is a simplified example. The well variable X_1 has only two categories and we code a dummy variable as 1 for the deep wells. The model becomes: $\hat{Y}_i = b_0 + b_1 \cdot X_{i1}$
  Relation to the simple *t*-test for difference in means. This leads to ANOVA. (see HAM p 86)

Type	$\bar{Y}$	s_Y	n
Shallow wells ($X_1 = 0$)	3.78	1.73	10
Deep wells ($X_1 = 1$)	3.07	1.26	42
Both	3.21	1.37	52

For shallow wells we get $\hat{Y}_i = 3.78 - 0.7 \cdot (0) = 3.78$ and for deep wells we get

$\hat{Y}_i = 3.78 - 0.7 \cdot (1) = 3.08$

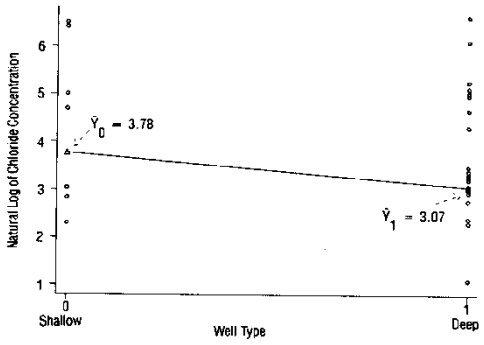

Figure 3.3 Regression of log chloride concentration on a dummy variable for well type.

- Different intercepts: Distance and intercept dummy

$\hat{Y}_i = 4.21 - 0.7 \cdot X_{i1} - 0.09 \cdot X_{i2}$

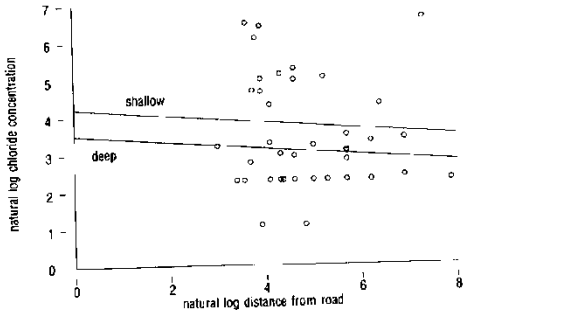

Figure 3.4 Regression of log chloride concentration on log distance from road and an *intercept* dummy variable for well type.

- Different slopes: distance and *interaction* between distance and well dummy variable

$$\hat{Y}_i = 3.67 - 0.03 \cdot X_{i2} - 0.08 \cdot X_{i1} \cdot X_{i2}$$

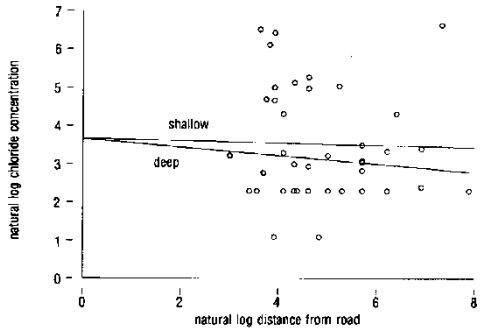

Figure 3.5 Regression of log chloride concentration on log distance from road and a *slope* dummy variable for well type.

- Different slopes and different intercepts: Combination of both previous models

$$\hat{Y}_i = 9.07 - 6.72 \cdot X_{i1} - 1.11 \cdot X_{i2} + 1.26 \cdot X_{i1} \cdot X_{i2}$$

For shallow wells $X_{i1} = 0$ the equation becomes

$$\hat{Y}_i = 9.07 - 1.11 \cdot X_{i2}$$

and for deep wells $X_{i1} = 1$ it comes

$$\hat{Y}_i = 2.35 + 0.15 \cdot X_{i2}$$

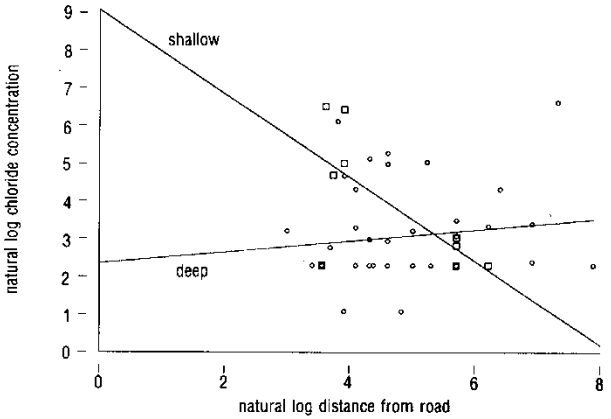

Figure 3.6 Regression of log chloride concentration on log distance from road, with slope and intercept dummy variables for well type.

All individual coefficients are significantly different from zero (see Table 3.4)

- A ***partial F-test*** can be conducted to test whether ***simultaneously*** the shallow/deep indicator variable and its interaction with the distance from the road is significant

$$F = \frac{(95.44 - 77.54)/2}{77.54/48} = 5.54$$

It's *p*-value at 2 and 48 degrees of freedom is 0.007. So both effects are jointly relevant.

- Each model could have been estimated independently. However, the ***simultaneous*** *perspective* has several advantages:

(1) Nested models allows for partial testing of whether we have different regression regimes in the first place?
(2) We can work with larger degrees of freedom because all observations and not just a subset associated with a particular factor level are included in the model.

- Problems of approach: There is the possibility that increased multicollinearity within the model increases.
    - Correlation between the *indicator* variable with the *overall intercept*.
    - Correlation of the *interaction* term with the *main-effects* of its parent variables.
- Note for ***sets of indicator variables*** based on more than two categories:
  [a] Insignificant (measured by their *t*-value) indicator variable levels in a set of indicator variables for a categorical variable should be ***kept in the model***, because
    - the significance (level parameter difference from zero) of individual indicator variables depends on which ***arbitrarily level*** has been dropped from the multi-level indicator variable.

However, different factor levels with identical parameters can be merged into a new category

- There are other ***coding specifications*** for the indicator variables than just the binary [0/1] coding scheme, which change the estimated regression parameters of the included factor levels but not the predicted value of the dependent variable.

  $\Rightarrow$ This leads to different significance levels for the individual categories

[b] The set of indicator variables is related to ***one factor***. Thus the whole categorical factor with its different levels needs to be tested by a ***partial* F-*test***.
That is, a model with all indicator variables versus a model without any indicator variables.
$\Rightarrow$ Thus excluding the whole categorical factor and not just individual levels.

- Note: Factor levels can be pooled together if their impact is the same. This requires a theoretical justification and should not just be a data-driven decision.

OTHER TOPICS: THE USE OF INDICATOR VARIABLE AND THE ANALYSIS OF VARIANCE (HAMILTON PP 92-101)

- ***One-way Analysis of Variance.*** You should be able to follow that example based on the radon exposure and lung cancer data.
  It can be formulated as regression analysis with a set of dummy variables for the individual factors.
- Alternative ***coding schemes*** *for categorical factors* and their properties
- ***Two-way Analysis of Variance***. It can be formulated as regression analysis with several dummy variables.
- ***Balanced designs***. Each sub-category has the same number of case which leads to uncorrelated factor levels in the Two-way ANOVA model.
- ***Analysis of Covariance***. Can be formulated as regression with categorical factors and metric variables.