Linear regression model

- The sample regression model is:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \dots + \underset{\overset{\downarrow}{?}}{\hat{\beta}_k} X_{ki} + \underset{\overset{\downarrow}{?}}{\hat{u}_i}$$

$$Y_i = \underset{\overset{\downarrow}{?}}{\hat{Y}_i} + \hat{u}_i$$

- What's the meaning of these estimators?
- Method of Ordinary Least Squares (OLS): It minimizes residual sum of squares (RSS):

$$min \sum_i \hat{u}_i^{\,2} \rightarrow min_{\hat{\beta}} \hat{u}'\hat{u} \rightarrow \hat{\beta} = (X'X)^{-1} X'Y$$

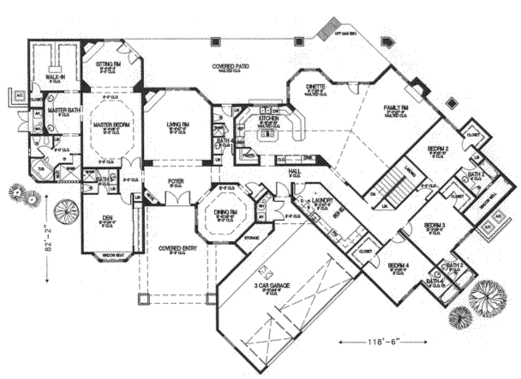
  - How do we obtain this?

Linear regression model

- Simple example:

$$house\_price(in\ thousands)_i = \beta_0 + \beta_1 bedrooms_i + \beta_2 sq\_fti + u_i$$

$$\widehat{house\_price} = -19.3 + 15.2 \times bedrooms + 0.13 \times sq\_ft$$



- What's the meaning of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$?

Linear regression model: Assumptions (again, review)

1) Model is *linear in the parameters*
2) Random sampling
3) No multicollinearity, or no perfect linear relationships among the X variables
4) Given X, the expected value of the error term is zero
   $E(u_i|X) = 0$
5) Homoscedastic, or constant, variance of u_i, or $var(u_i|X) = \sigma^2$

- What's the meaning of these assumptions?

(we will come back to these assumptions)

Linear regression model: Gauss-Markov Theorem

- On the basis of assumptions, the OLS method gives the Best Linear Unbiased Estimators (BLUE)
  1) Estimators are ***linear*** functions of the dependent variable Y.
  2) The estimators are ***unbiased***; in repeated applications of the method, the estimators approach their true values.
  3) In the class of linear estimators, OLS estimators have **minimum variance**; i.e., they are **efficient**, or the "***best***" estimators.

Assumptions violations

- Implications of the violations of the following assumptions (as review, as you saw this in the previous course), solutions, and applications
  - Independence of error terms - Autocorrelation
  - No perfect collinearity vs. Multicolllinearity
  - Variance of u_i does not depend on the value of x_i - Heteroskedasticity
  - $Cov(X_i, u_i) = 0$
    - Omitted variables bias (we will cover several solutions about this one in this course)
    - Also, measurement error and simultaneous equation models
- What are the consequences and potential solutions?

Omitted variable bias

- What if we say that people who play golf are more prone to heart disease and arthritis? (example from "Naked Statistics)
  - Are we missing something?
- Assume that the true model is given by

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \rightarrow \hat{\beta}_1$$

- But you estimate: $y = \beta_0 + \beta_1 x_1 + u \rightarrow \tilde{\beta}_1$
- Remember that u and x must be uncorrelated. Show that

$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1$, where $\tilde{\delta}_1$ is the slope coefficient for regressing x_2 on x_1

- This means that $\tilde{\beta}_1$ is biased. Let's see an example.

Omitted variable bias: direction of bias

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

- When the bias is equal to 0?
- So, adding more variables is the solution? No. Why?

Example

- Restaurant new location for a established restaurant chain based on volume of customers (Y), using competition (N), population (P) and average household income (I) from nearby places (from Stundenmund, 2016)

```
--------------------------------------------
                     est1            est2
                      b/t             b/t
--------------------------------------------
N              -9074.674***   -1487.344
                  (-4.42)      (-0.84)
P                  0.355***
                   (4.88)
I                  1.288*          2.322**
                   (2.37)         (3.50)
Constant      102192.428***  84438.590***
                   (7.98)         (5.19)
--------------------------------------------
N                  33.000         33.000
r2                  0.618          0.305
r2_a                0.579          0.258
--------------------------------------------
* p<0.05, ** p<0.01, *** p<0.001
```

- What would you expect about Corr(N,P)?
  - Actually, we know it in this case
- So, what would it be the direction of the bias?
- Is it consistent with the results?

Model specification errors

- One of the assumptions of the classical linear regression (CLRM) is that the model is correctly specified. For example:
  - The model does not exclude any “core” variables
  - The model does not include superfluous variables
  - The functional form of the model is suitably chosen

- If you have a model (called “restricted”), and add (a relevant) variable, how can you compare this new “unrestricted” model with the original one?
  - Ramsey RESET test (about functional form)

Model specification: Choosing “the right” variables

- You may have many potential covariates – and many ways to specify the right hand side
- Think about the theoretical and economic relevance of each variable
  - Researchers may end up using many predictors (including exponents and interactions)
  - Don’t use variables just because they are available
  - Some variables may be statistically significant due to random chance – they may also affect other predictors (think about multicollinearity)
    - (sadly a real example) Does a bad grade in chemistry predict that you will work in a big company?