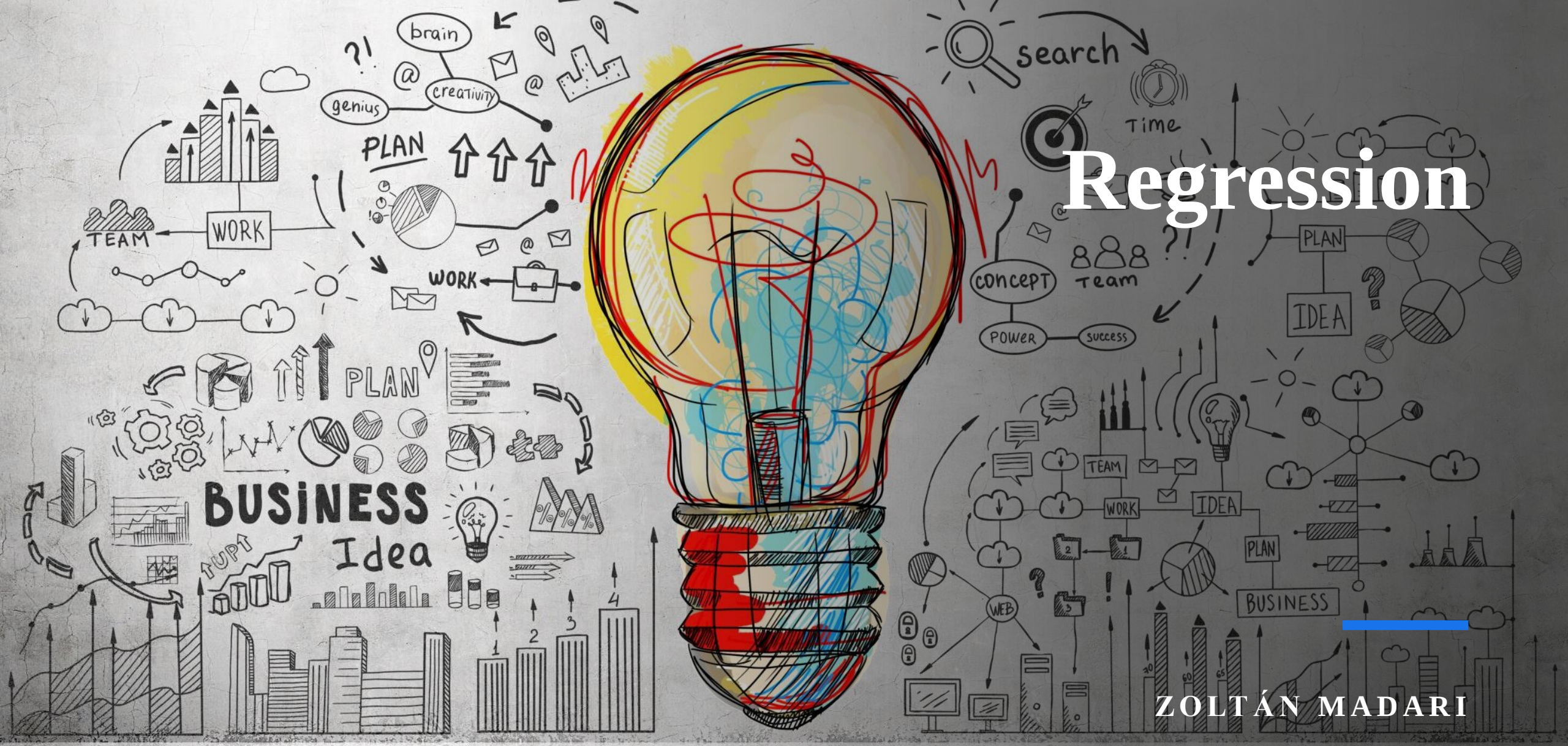


Regression

ZOLTÁN MADARI

Empirical analysis

Econometrics: The application of mathematical statistics to empirical research in economics

- Estimating relationships between variables
- Testing economic theories in real life
- Forecasting trends
- Evaluating effects of government policies, monetary and fiscal decisions

The applied method in most cases is linear regression.

Example:

1. **Formulating the question of interest–** How does the number of school years effect one's wage?
2. **Structural model**

$$Wage \sim f\,(level\ of\ educ, quality\ of\ educ\ , work\ experience, sex, abilities)$$

3. **Econometric model**– regression equation

$$Wage = \beta_0 + \beta_1 lev.of\ educ. + \beta_2 work\ exp. + \beta_3 age + u$$

4. **Research hypothesis** $\beta_1 > 0$

Covariance and correlation

Covariance is a measure of the joint variability of two random variables. Small values correspond with weak, greater values with strong connection between variables. Nonnormalized; by normalization we get the correlation coefficient.

$$Cov(X,Y) = \mathbb{E}\big(\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\big)$$

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

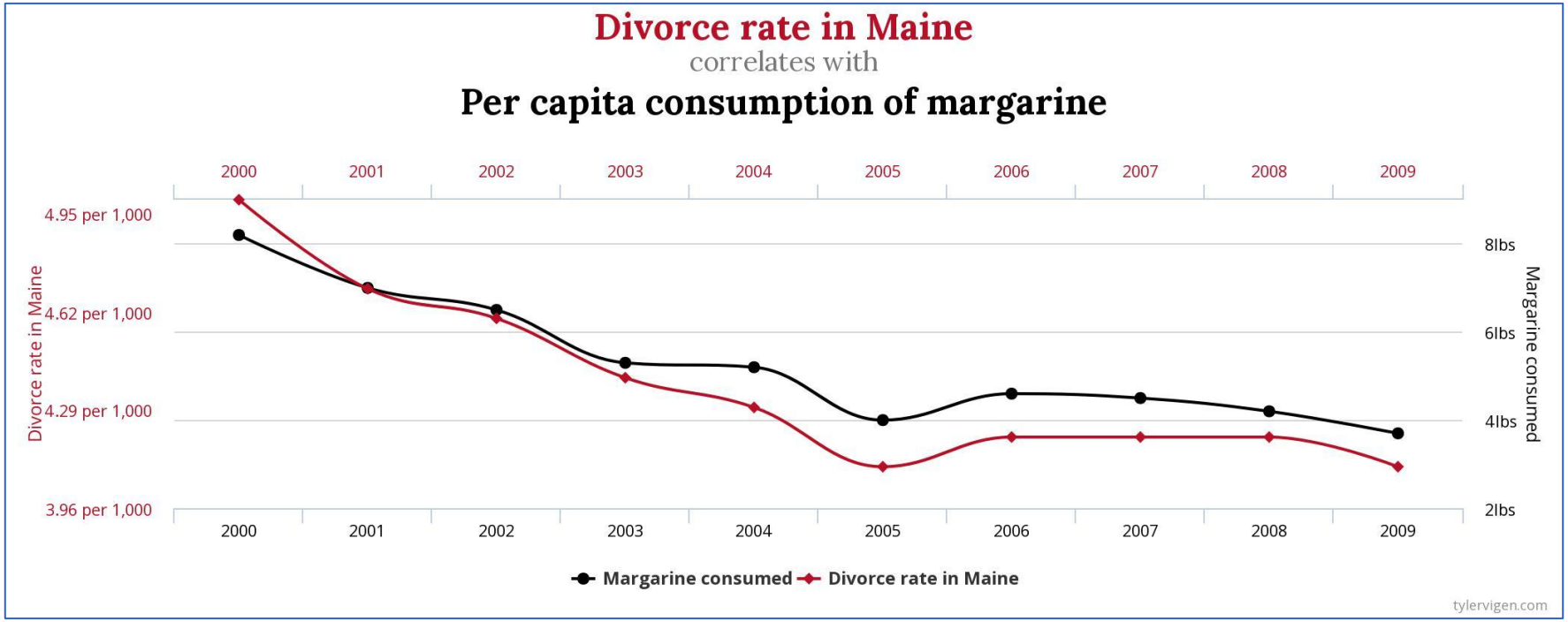
Correlation is a statistical measure that expresses the direction and extent to which two variables are linearly related.

$$Corr(X,Y) = \frac{\mathbb{E}\big(\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\big)}{\mathbb{D}(X) \cdot \mathbb{D}(Y)} = \frac{Cov(X,Y)}{\mathbb{D}(X) \cdot \mathbb{D}(Y)}$$

$$corr(x,\ y) = \rho_{x,y} = \frac{\mathrm{cov}(x,\ \mathrm{y})}{\sigma_x \cdot \sigma_y}$$

Correlation ≠ causality

A correlation between variables does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.



https://www.tylervigen.com/spurious-correlations

The problem of confounding

Confounding:

The deviation of a variable coheres with the deviation of the examined aspect of classification meanwhile also affecting the dependent variable.

We usually call this as confounder variable. It can distort the measured effect of explanatory variable.

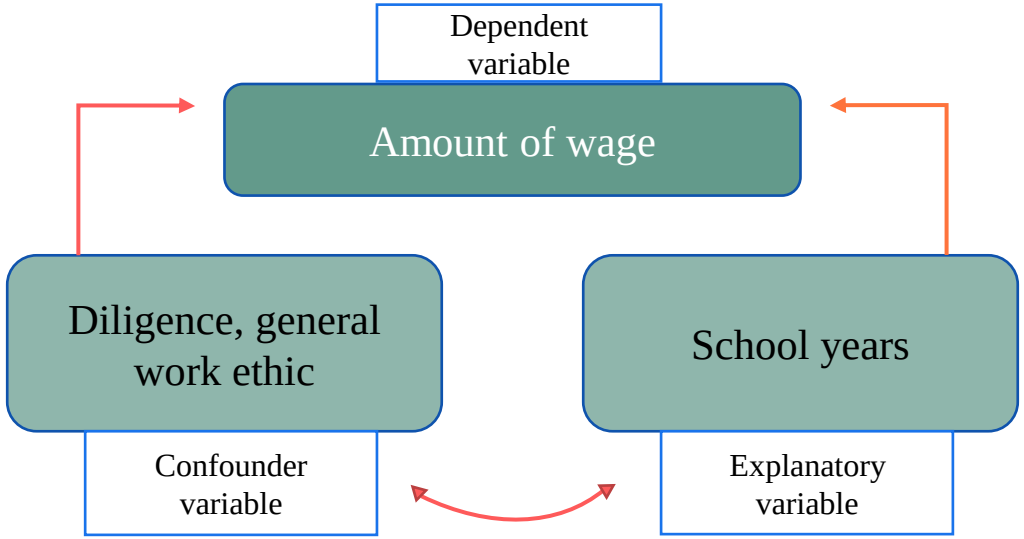
How can we avoid it?

Goal: the examined groups should only differ by the chosen aspect (which we want to measure the effect of) .

- Experiment based analysis
- Observation based analysis – by deploying econometric concepts

Example:

„More school years come with greater wage”



Bivariate regression I.

Empirical regression is a technique in non−parametric regression. The key idea is to estimate the joint density of X, Y, denotes p(X,Y). From this joint density one can derive conditional p(Y|X), and conditional mean E(Y|X). This E(Y|X) is empirical regression.

Analytical regression function is defining the stochastic relationship between X and Y by a function f(X)

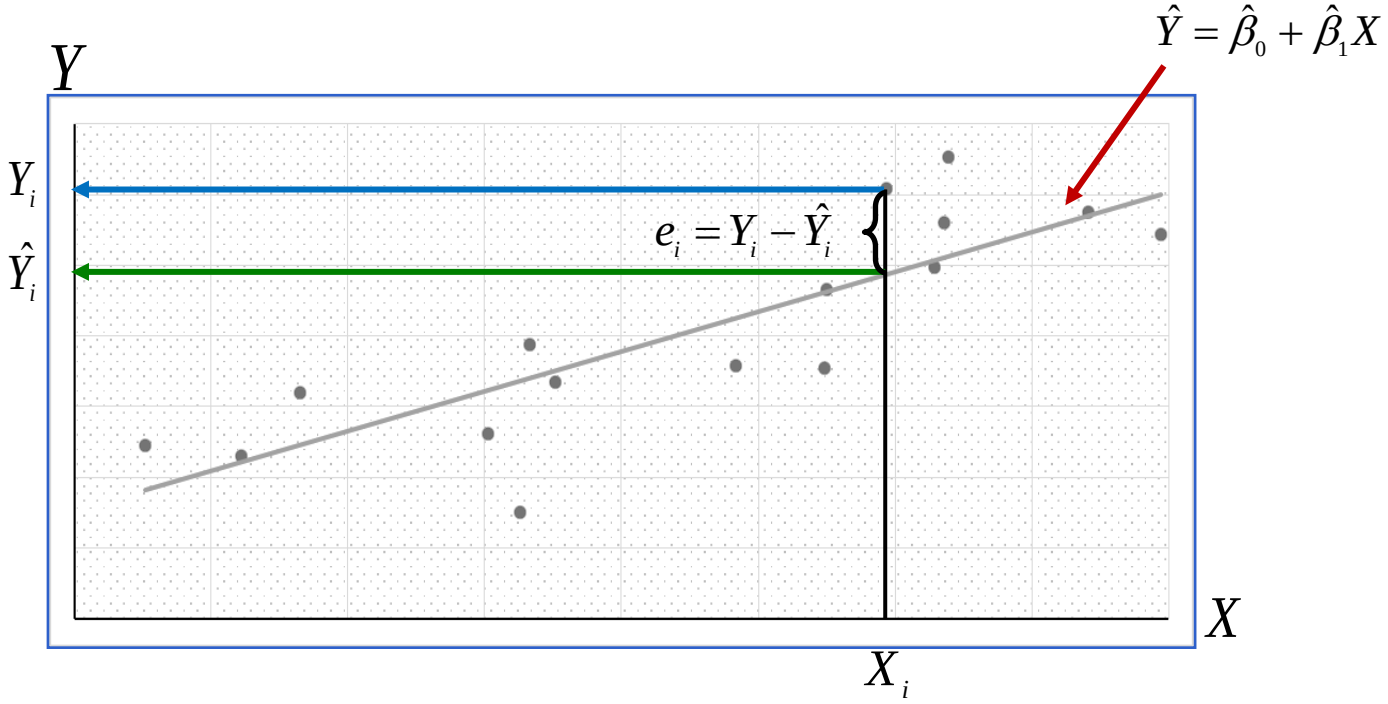
$$f(X) = \beta_0 + \beta_1 X + \varepsilon$$

Equation of bivariate analytical regression

Estimated parameters $\hat{\beta}_0, \hat{\beta}_1$

The given parameters are estimated using the ordinary least squares – OLS – method.

Bivariate regression II.



Interpretation of estimated parameters

$\hat{\beta}_o$

Intercept – if the explanatory variable takes the value 0, this will be the value of the dependent variable according to the model

$\hat{\beta}_1$

The average effect in the estimated dependent variable (in the examined range) is caused by a unit change in the value of the explanatory variable

$$g = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

error minimization, extreme value problem

Model assumptions

1. **Linearity**
2. **No exact multicollinearity**
3. **Strong or strict exogeneity**
4. **Homoscedasticity**
5. **No autocorrelation**

In case of IID sample

1. $f(X) = \beta_0 + \beta_1 X + \varepsilon$

2. The data matrix has full column rank.

   Variables cannot be written as a linear combination of each other

3. $\mathbb{E}(\varepsilon_i | X_i) = 0$

   The errors are independent of the explanatory variables

4. $\mathbb{D}^2(\varepsilon_i | X) = \sigma^2$

   The standard deviation of errors for different observations is constant

5. The errors for different observations are uncorrelated

An indicator of „goodness” of the model – R^2

- It measures the fit of the regression line and the explanatory power of the model
- It can take values between 0 and 1 - it can also be written as a percentage

The regression model can be located in this interval

Null model with only a constant term – estimates the mean as the outcome variable

Perfect model

An indicator of „goodness” of the model – R^2

$$\sum\left(Y_i-\bar{Y}\right)^2=\sum\left(\hat{Y}_i-\bar{Y}\right)^2+\sum\left(Y_i-\hat{Y}_i\right)^2$$

SST	SSE	SSR
Total sum of squares	Explained sum of squares	Sum of squared residuals

R^2 or coefficient of determination

$$R^2=\frac{SSE}{SST}=1-\frac{SSR}{SST}$$

Only in the bivariate case the determination coefficient equals to the square of the correlation coefficient

$$R^2=\frac{SSE}{SST}=\frac{Var(\hat{Y})}{Var(Y)}=\frac{Var(\beta_0+\beta_1 X)}{Var(Y)}=$$

$$=\beta_1^2\frac{Var(X)}{Var(Y)}=\frac{Cov(X,Y)^2}{Var(X)^2}\cdot\frac{Var(X)}{Var(Y)}=$$

$$\frac{Cov(X,Y)^2}{Var(X)Var(Y)}=Corr(X,Y)^2$$

Inference in regression

1. **Setting up hypothesis**
2. **Selection of the test statistic**
3. **Specifying critical values and/or p-value**
4. **Making the decision at a given significance level**

Separate testing of parameters

$H_0: \beta_1 = 0$

The parameter of a given explanatory variable is 0 -> it has no significant explanatory power

$H_1: \beta_1 \neq 0$

The parameter of a given explanatory variable differs from 0 -> has significant explanatory power

When testing the significance of the estimated parameters separately, a t-test is used

Testing the model as a whole

$$H_0: \ \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

The combined explanatory power of the model's explanatory variables is 0 -> the model is completely wrong

$$H_1: \ \exists \beta_j \neq 0$$

There is at least one coefficient which differs from 0, i.e. its explanatory power is significant

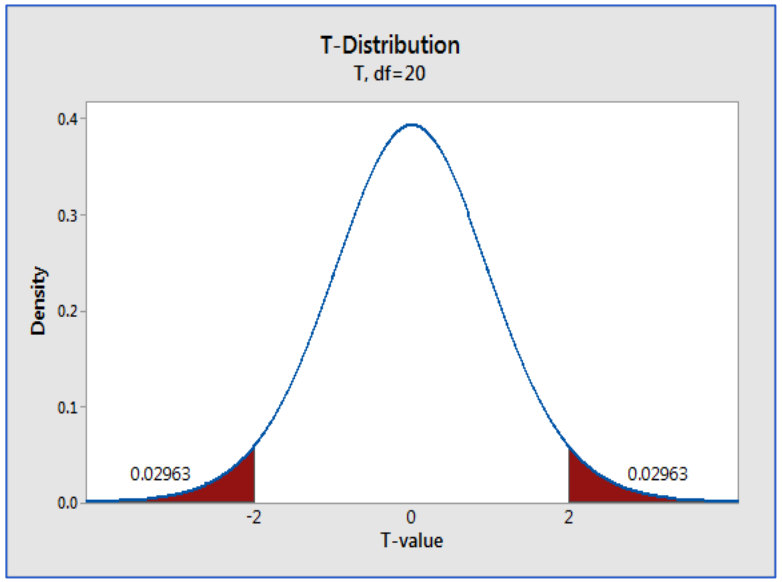
F-test is used when testing the significance of the entire model (analysis of variance).

T-test

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Hypotheses



$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_j}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

Test statistic

$$t_{1-\frac{\alpha}{2}}(n-p)$$

Test statistic distribution

Rejection and nonrejection regions

Multiple linear regression

Assumption: multivariate normal distribution

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \mathrm{u}$$

Where:

Y: dependent variable

$\boldsymbol{\beta_0}$: intercept

$\mathbf{X_i}$: explanatory variables

u: error term

Estimated value:

$$\hat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \widehat{\beta_2} X_2 + \cdots + \widehat{\beta_k} X_k$$

In matrix form:

$$Y = X\beta + \varepsilon$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & & X_{2p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & & X_{np} \end{bmatrix}$$

β estimation by OLS

Objective function:

$$\min_{b}[(y - Xb)'(y - Xb)]$$

In traditional form:

$$\min_{b_1 \ldots b_k} \sum_{i=1}^{n} \left(y_i - b_1 - \sum_{j=2}^{k} b_j x_{ij} \right)^2$$

Approximate value:

$$\hat{y} = Xb = X(X'X)^{-1}X'y$$

β interpretation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_j x_j + \ldots + \hat{\beta}_k x_k$$

Partial derivative

$$\frac{\partial \hat{Y}}{\partial x_i} = \beta_i$$

Interpretation of slope (β_i): for every 1-unit increase in the given explanatory variable, the dependent variable will increase by the beta coefficient value. (average effect)

Interpretation of intercept β_0 : if all explanatory variables are equal to zero, it is the **expected value of the dependent variable**

- Usually it is just a fitting parameter

Ceteris paribus: all other things being unchanged or constant

The content of the *u* error term

1. Omitted variables

2. Misspecified functional form

The true population model: $y = f(\mathbf{x}) + u$

The estimated model: $y = g(\mathbf{x}) + \varepsilon$, where $\varepsilon = f(\mathbf{x}) - g(\mathbf{x}) + u$

A simple example

The true population model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$

The estimated model: $y = \beta_0 + \beta_1 x + \varepsilon$, where $\varepsilon = \beta_2 x^2 + u$

3. Measurement errors in the variables

The true population model: $y = \beta_0 + \beta_1 x + u$

The estimated model: $y = \beta_0 + \beta_1 (x + error) + \varepsilon$, where $\varepsilon = u - \beta_1 \cdot error$

4. Random variation of *y* around its expected value (the random error)

Global F-test

We test the whole model

$H0: \beta_1 = \beta_2 = \beta_k = 0$ None of the explanatory variables affects significantly the dependent variable

$H1: \exists i, \beta_i \neq 0$ There is at least one significant variable in the model

Test function

$$F = \frac{SSE/k}{SSR/(n-p)}$$

Where,
k: number of explanatory variables
n: number of obsevations
p: number of estimated parameters

It follows F-distribution

Interval estimation for regression parameter

The 95% confidence interval of β_j :

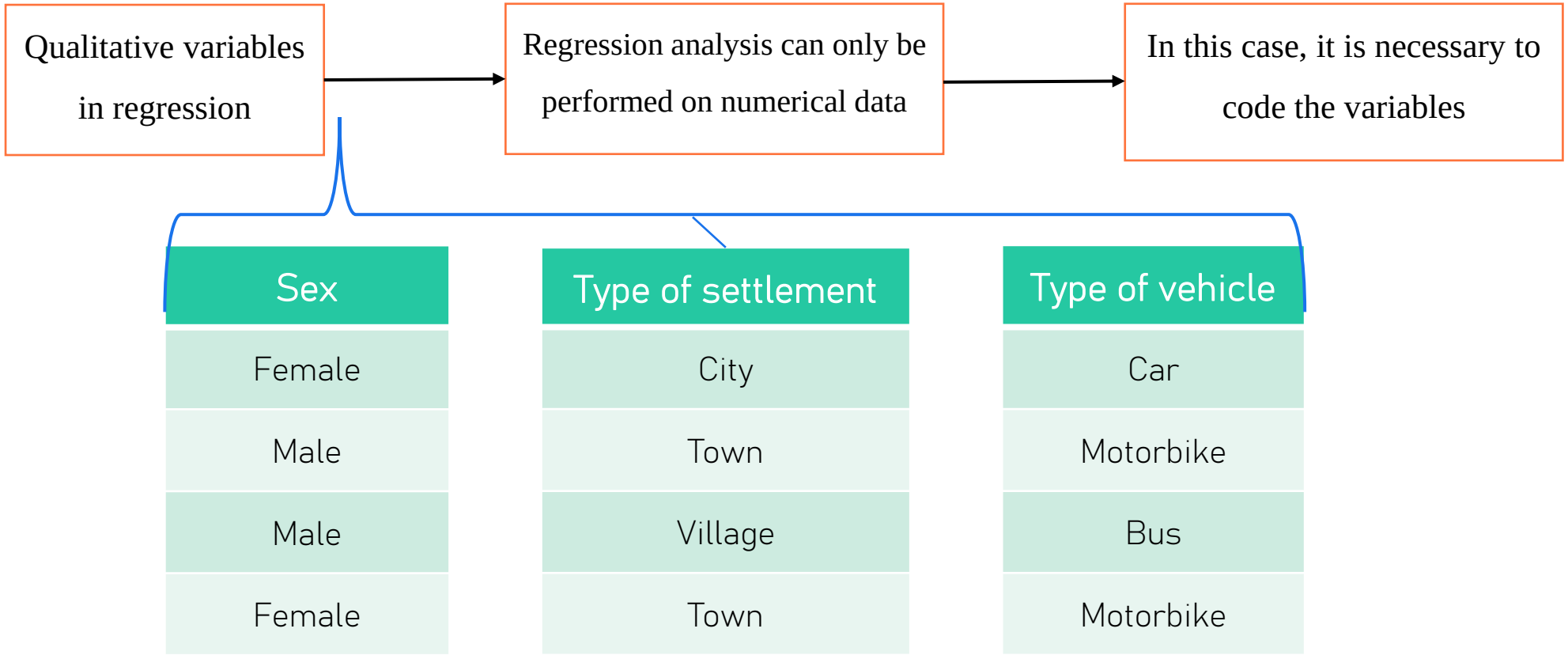
$$CI95\_\beta_j = \hat{\beta}_j \pm t_{0.975} \cdot se(\hat{\beta}_j) = \left[\beta_j^L,\ \beta_j^U \right]$$

$t_{0.975}$ is a multiplier value from t-distribution

se is the standard error of given parameter

$t_{0.975}$ * se is the margin of error

Categorical/qualitative variables



Dummy coding

The categorical variables can be coded so that they take the values 0 and 1

Sex		Female	Male
Female	→	1	0
Male	→	0	1
Male	→	0	1
Female	→	1	0

The data matrix will have as many new columns as the number of outcomes in the original variable

In the case of N parameter variants, the Nth column is redundant, the coding can be carried out with N-1 columns

Re: Model assumptions ~ no exact multicollinearity

Including a dummy variable in the regression

Column rank of data matrix

One of the main steps of the OLS method is to invert the matrix -> this can only be done with full column rank -> i.e. if the variables cannot be written as a linear combination of each other.

For this reason, in addition to dummy coding, all categories can never be included in the model at the same time, we must define a **reference cathegory** or omit the constant.

With constant:

$$\begin{pmatrix} 1 & A \\ 1 & B \\ 1 & C \end{pmatrix} \sim \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

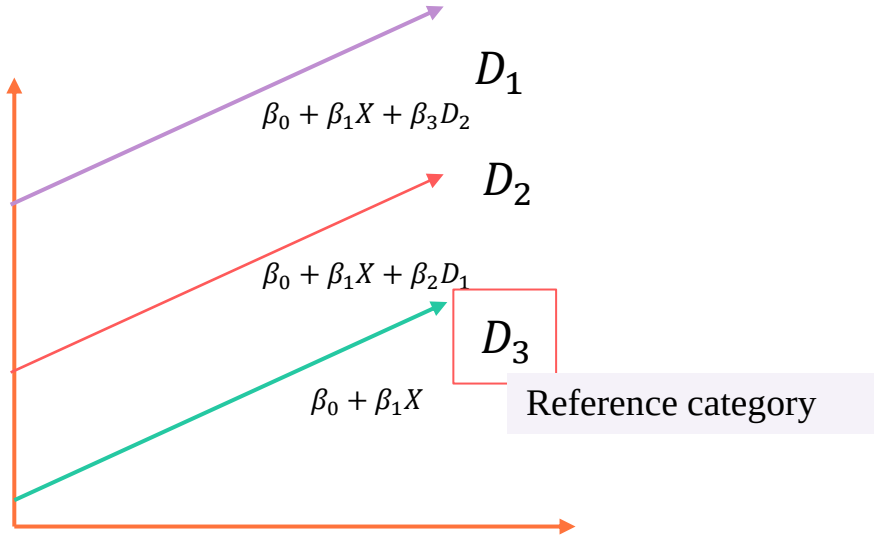
Without constant:

$$\begin{pmatrix} A \\ B \\ C \end{pmatrix} \sim \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Interpreting dummy variables

Dummy variable in the regression equation

$$f(X) = \beta_0 + \beta_1 X + \beta_2 D_1 + \beta_3 D_2 + \varepsilon$$



Dummy variables modify the intercept of the regression line for different categories

Logic of model building

By **increasing the number of explanatory variable** in the model, the value of the coefficient of determination (R^2) **certainly won't decrease**.

Based on this, the best decision to use **all potential explanatory variables**.

- In reality it is not sure!
- High R^2 means good fit on the sample, but we want to describe **the population**

TARGET

To determine the minimum range of variables that have **a meaningful, statistically measurable effect** on the dependent variable.

Conclusion

The model should contain reasonable number of explanatory variables, which could significantly explain the variance of dependent variable.

R^2 and adjusted R^2

R^2

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

- Explanatory power of the model
- PRE-indicator
- Limits:

$$0 \leq R^2 \leq 1$$

Adjusted R^2

$$1 - (1 - R^2)\frac{n-1}{n-p}$$

- R^2 will increase if we add a new variable to the model.
- Adjusted R^2 punishes the unneccesary explanatory variables.
- It could be negative.
- Adjusted R^2 is already suitable for comparing model with different numbers of independent variables

Information criteria (IC)

Akaike: $AIC = \frac{SSR}{n} e^{\frac{2k}{n}}$

Schwarz: $BIC = \frac{SSR}{n} n^{\frac{k}{n}}$

Hannan-Quinn: $HQC = \frac{SSR}{n} (\ln n)^{\frac{2k}{n}}$

They penalize **the large number of explanatory variables** and **the large error** at the same time, try to find balance between the SSR and number of parameters.

Where,

n, number of observations

k, number of parameters

- They are built on a completely different principle (based on information theory) than the adjusted R^2
- **Error indicators,** therefore the goal**: minimization**
- Not only nested models can be compared

Wald test

U: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

R: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

$\boldsymbol{H0}: \beta_{p+1} = \beta_{p+2} = \beta_{p+q} = 0 \; or \; R_U^2 = R_R^2$

$\boldsymbol{H1}: \exists i, \beta_i \neq 0 \; or \; R_U^2 > R_R^2$

$$\frac{(R_{UR}^2 - R_R^2)/q}{(1 - R_{UR}^2)/(n - p_{UR})} \square F_{(q, n-p_{UR})}$$

Where
n: number of observations
q: number of omitted variables (restrictions)
p: number of remaining variables including constant

Wald-test for omitted variables

- We decide between two models, between an unrestricted (U) and a restricted (R) one
- **Nested modelselection**: all variables of the narrower model are included in the wider model

H0: the specified m variables **do not even have a significant explanatory power jointly**– they can be omitted from the regression model.

$\boldsymbol{F(q, n - pUR)}$

Lagrange multiplier test

U: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

R: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

$\boldsymbol{H0}: \beta_{p+1} = \beta_{p+2} = \beta_{p+q} = 0 \; or \; R_U^2 = R_R^2$

$H1: \exists i, \beta_i \neq 0 \; \boldsymbol{or} \; \boldsymbol{R_U^2 > R_R^2}$

- Idea: estimate the restricted model and calculate the estimated residuals based on it
- If H0 exists, these residuals can not be significantly explained either by the variables of the restricted model (consequence of OLS) or by the possible q variables (consequence of H0)

$$LM = n \cdot R_{\hat{u}}^2 \; \square \; Chi^2_{DF:q}$$

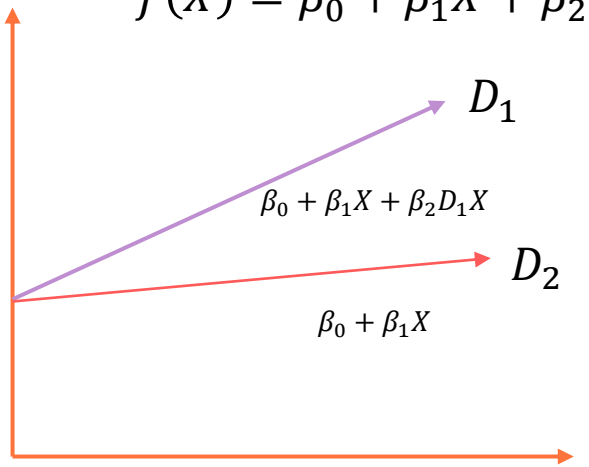
Interaction in regression

Interaction

Membership in a given group often affects the effect of one of the explanatory variables -> this can be included in the model by including interaction term

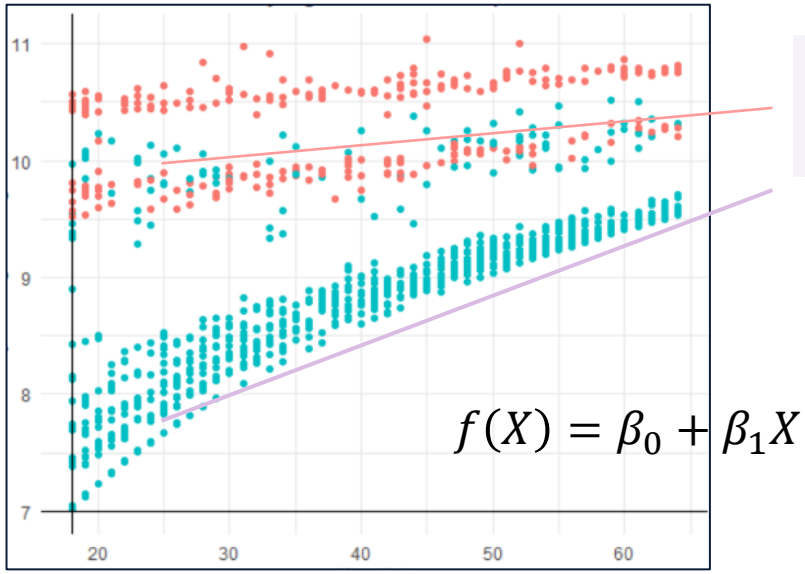
The deviation of the intercept and the effect can also be included within one model.

$$f(X) = \beta_0 + \beta_1 X + \beta_2 D_1 X + \varepsilon$$



Model with interaction term

Example: relation of insurance fee and age for smokers and nonsmokers



Model with interaction term and dummy variable

$$f(X) = \beta_0 + \beta_1 X + \beta_2 D_1 + \beta_3 D_1 X + \varepsilon$$

Partial derivative

Marginal effect

Marginal effects tells us how a dependent variable (outcome) changes when a specific independent variable (explanatory variable) changes. Other covariates are assumed to be held constant.

Typical simplification: how much does the dependent variable change as a result of increasing the explanatory variable by one unit

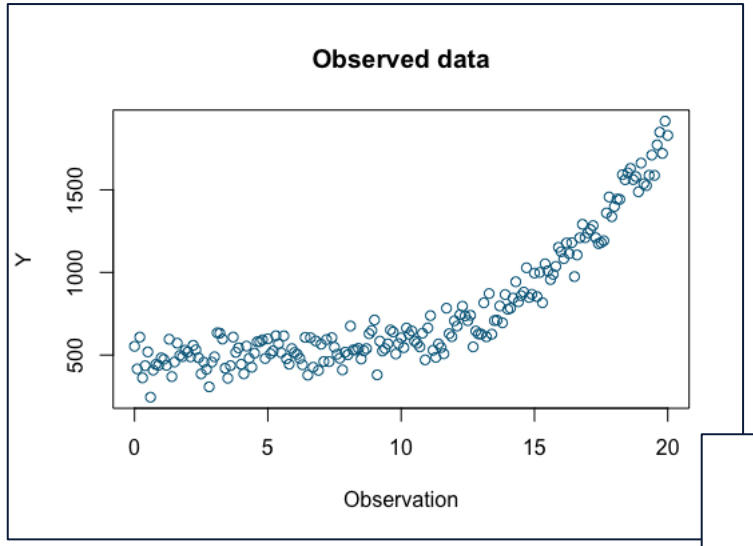
However, in the case of more complex models (e.g. including interaction terms), we define it according to the precise definition:

$$\frac{\partial Y}{\partial X_j}$$

The partial derivative of the outcome variable according to the jth explanatory variable

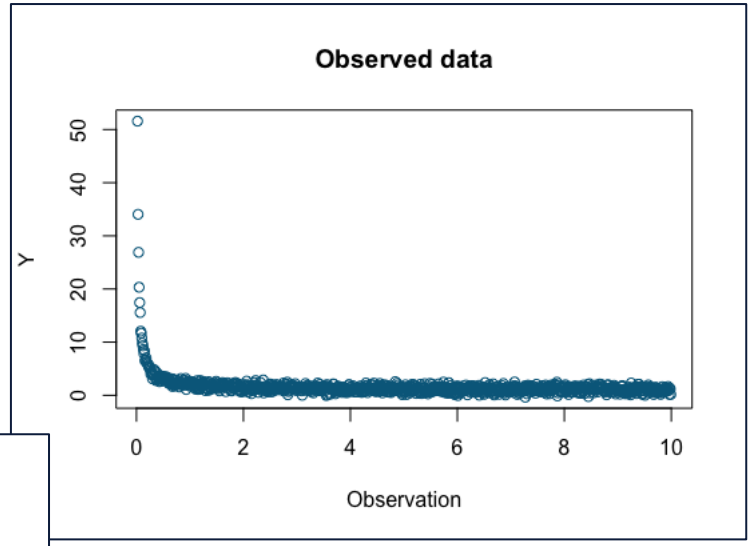
Nonlinearity

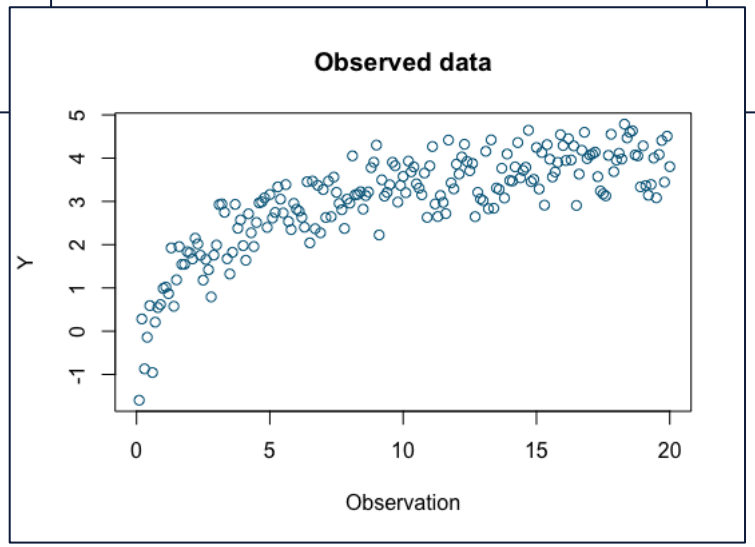
Reality is not always linear



Re: Model assumptions ~ linearity

We are not able to describe the given phenomenons by a simple linear line.





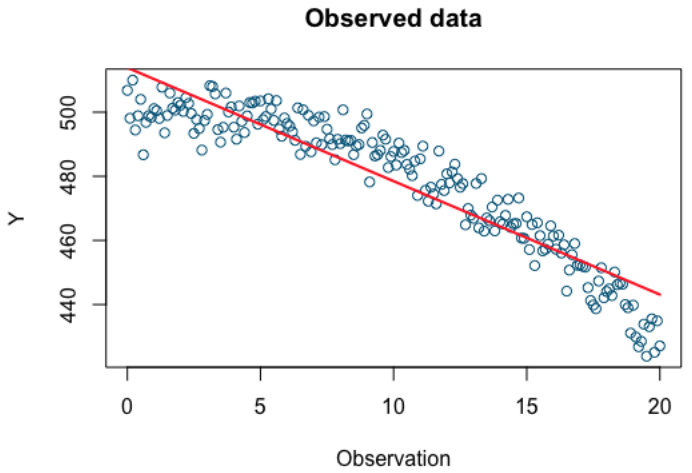
Quadratic terms

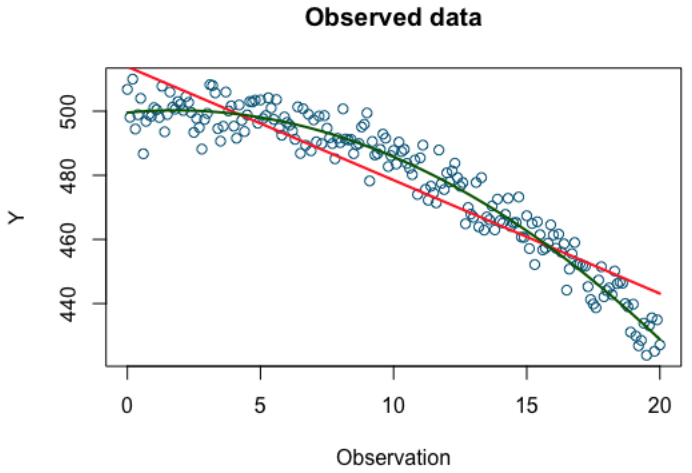
Add the quadratic form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Partial derivative – marginal effect of X

$$\frac{\partial Y}{\partial X} = \beta_1 + 2 * \beta_2 X$$





Four cases of logarithmic transformations

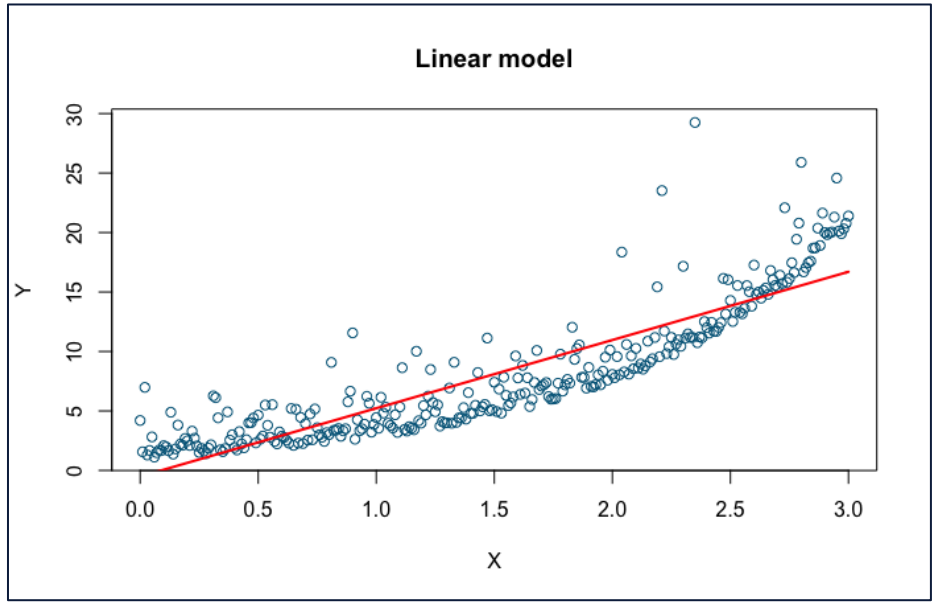
Resolving linearity with logarithmic tranformations

	X	**logX**
Y	Linear $\widehat{Y}_i = \alpha + \beta X_i$	Linear-log $\widehat{Y}_i = \alpha + \beta \log X_i$
Log Y	Log-linear $\log\widehat{Y}_i = \alpha + \beta X_i$	Log-log $\log\widehat{Y}_i = \alpha + \beta \log X_i$

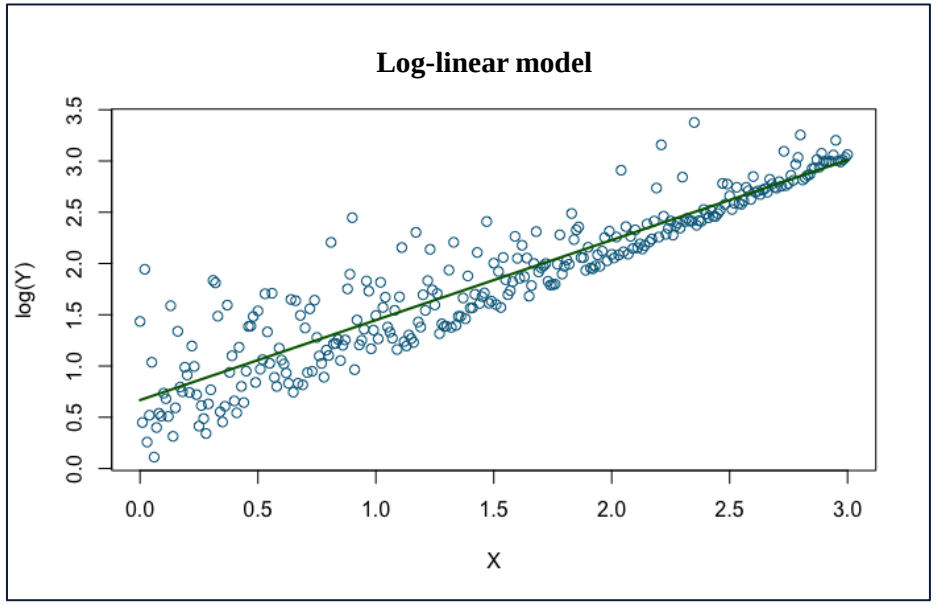
Taking the logarithm of a variable is useful to linearize the relation between two variables.
It is also a useful tool to handle skewness in case of the variables.

Log-linear model

Original population: $Y_i = e^{\alpha + \beta X_i + \varepsilon}$



$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + u$$

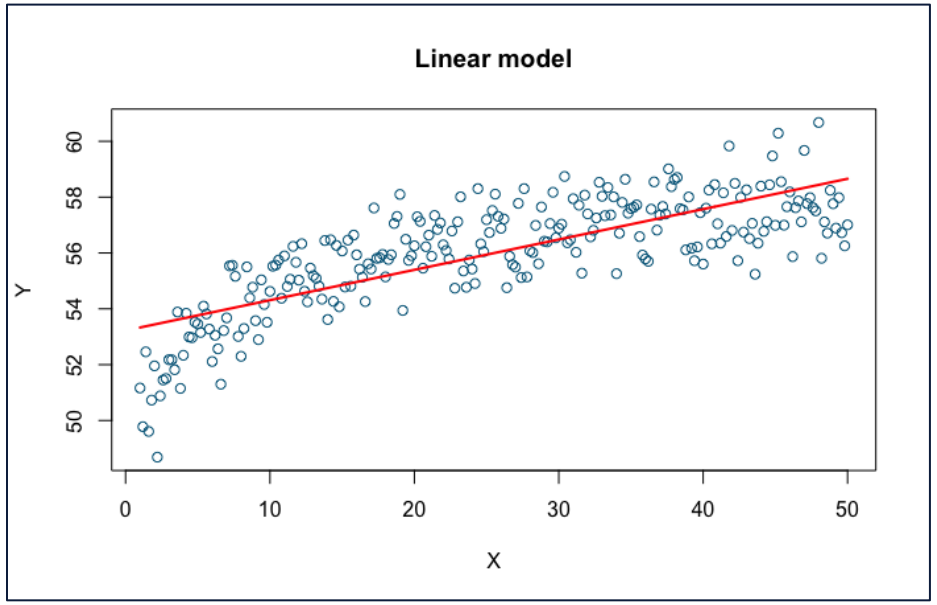


$$\log\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + u$$

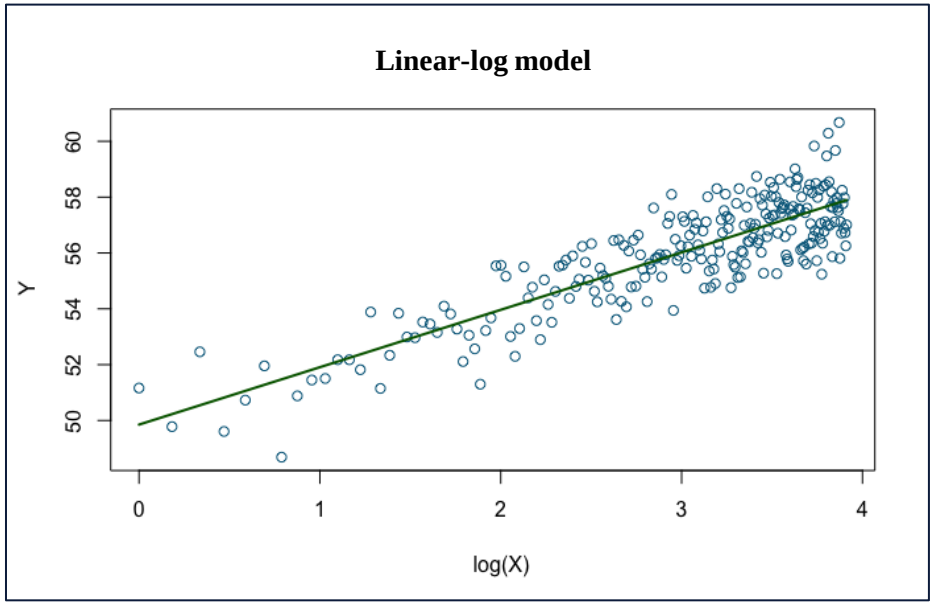
Interpretation: Each 1-unit increase in X multiplies the expected value of Y by e^{β}

Linear-log model

Original population: $Y_i = \alpha + \beta \log(X_i) + \varepsilon$



$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + u$

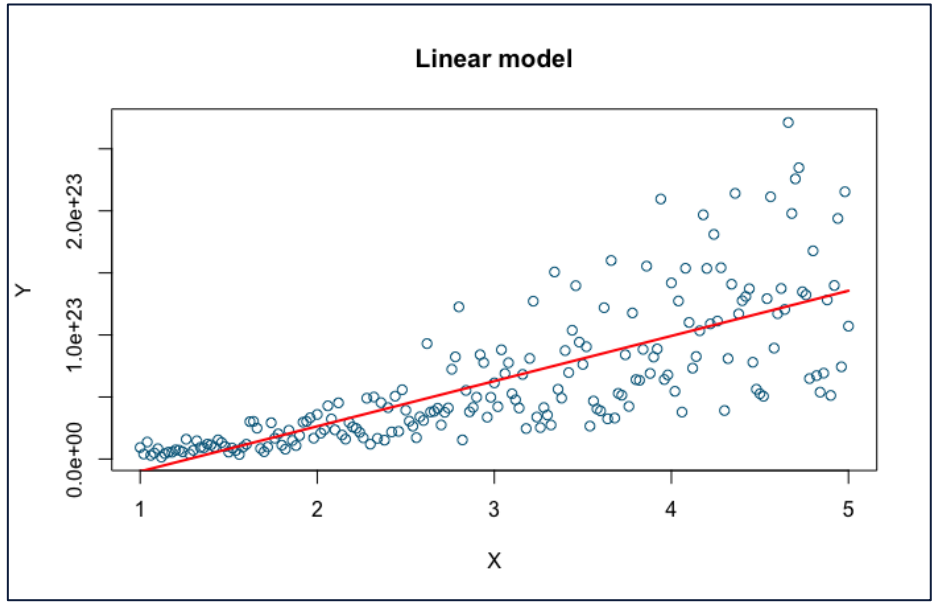


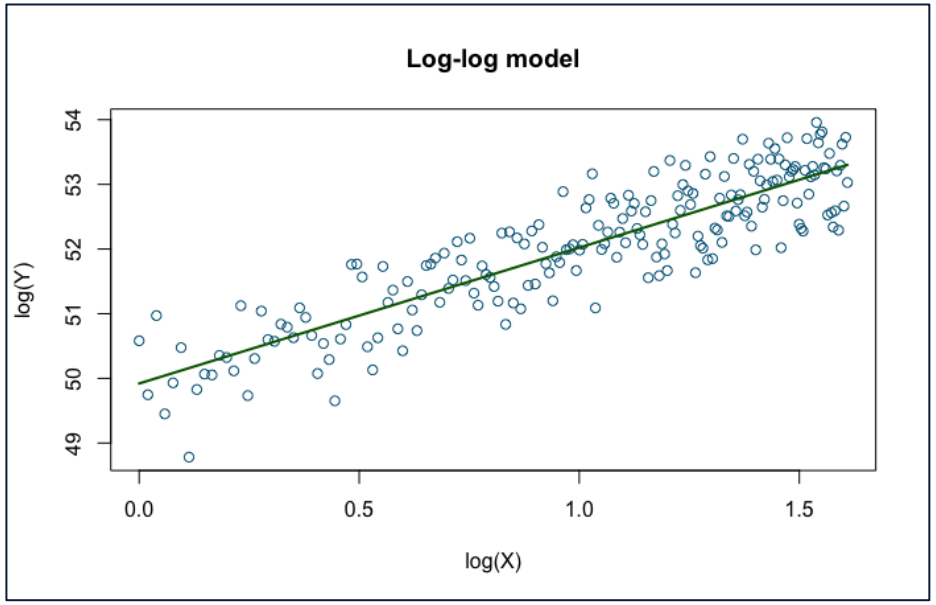
$\hat{Y}_i = \hat{\alpha} + \hat{\beta} \log(X_i) + u$

Interpretation: one percentage change in X results $\beta/100$ units change in Y

Log-log model

Original population: $Y_i = e^{\alpha+\beta\log(X_i)+\varepsilon}$





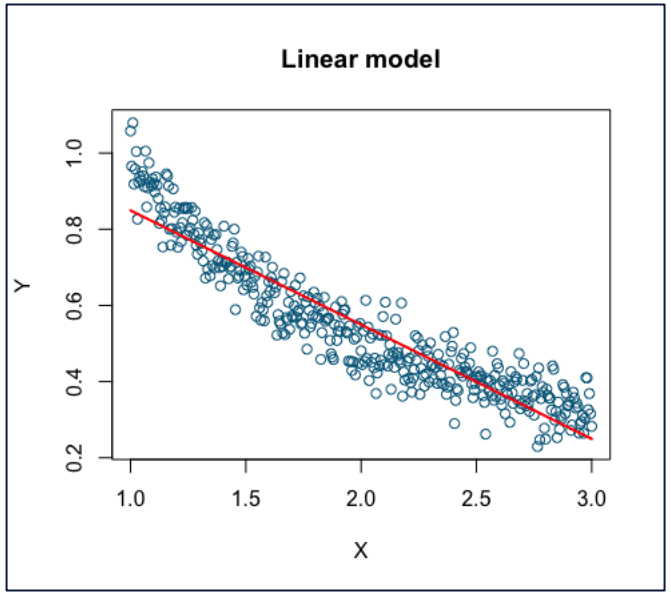
$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i + u$$

$$\log(\hat{Y}_i) = \hat{\alpha} + \hat{\beta}\log(X_i) + u$$

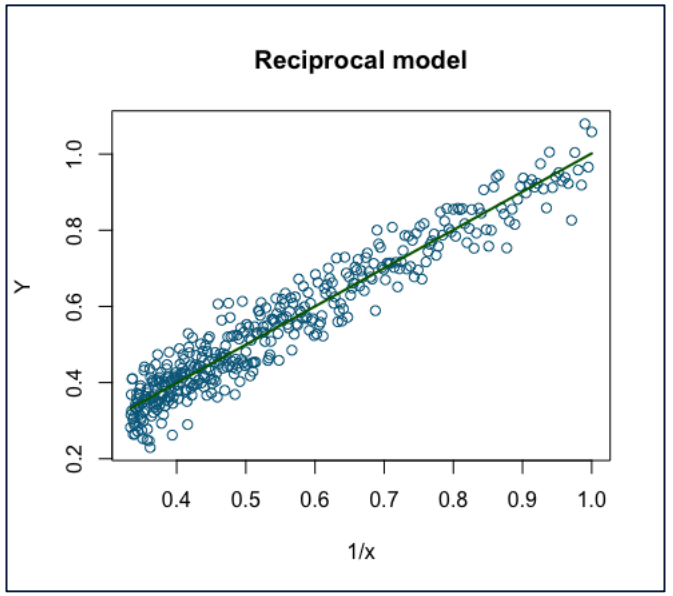
Interpretation: one percentage unit increase in X results β percentage change in Y

Reciprocal model

Original population: $Y_i = \alpha + \beta \frac{1}{X} + \varepsilon$



$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i + u$$



$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} \frac{1}{X_i} + u$$

Interpretation: The marginal effect of a small increase in X is $\frac{\partial Y}{\partial X} = -\frac{\beta}{X^2}$

Ramsey RESET test

The Ramsey Regression Equation Specification Error Test (RESET) test is **a general specification test for the linear regression mode**l. More specifically, it tests whether non-linear combinations of the fitted values help explain the response variable.

Advantage: It does not look for an answer to a specific specification question, but generally examines whether the specification is good;

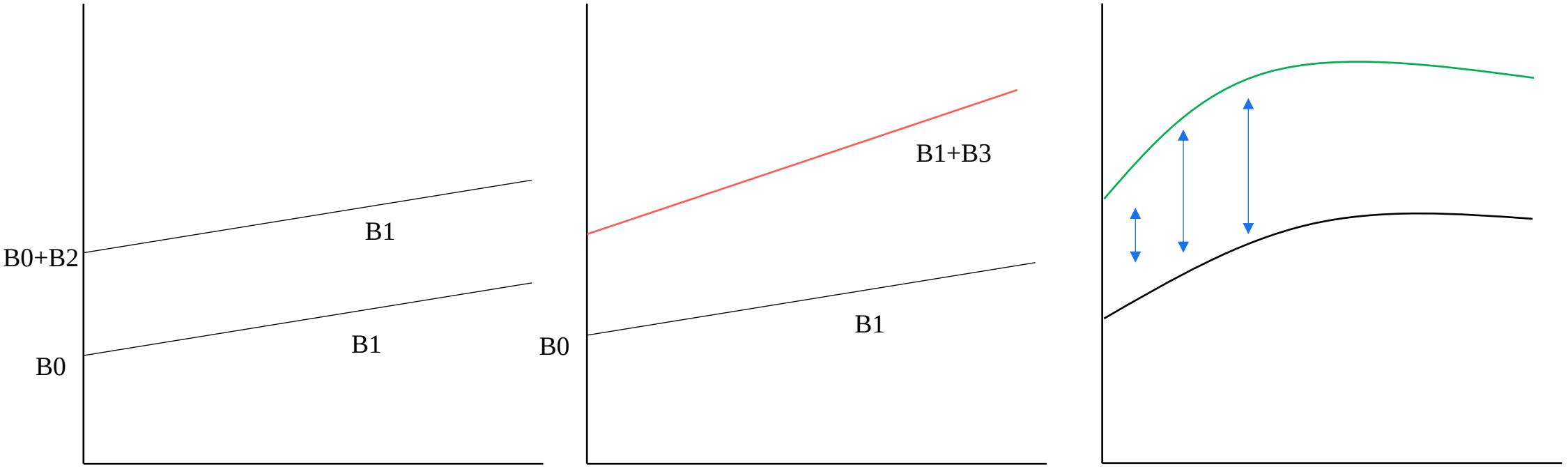
Disadvantage: if it gives a negative answer, the exact problem with the given specification remains unkonwn

$$Y = \beta'_0 + \beta'_1 X_1 + \beta'_2 X_2 + \ldots + \beta'_k X_k + \gamma_1 \widehat{Y}^2 + \gamma_2 \widehat{Y}^3 + \varepsilon'$$

$$H_0 : \gamma_1 = \gamma_2 = 0$$

Estimation of wage equation – is there wage discrimination?

Wage=B0+B1*Exp+B2*D_Male+B3*D_Male*Exp+B4*Exp^2+e



Marginal effect of experience in the last model: B1+B3*D_Male+2*B4*Exp
The effect of experience depends on the sex and the level of experience!!!