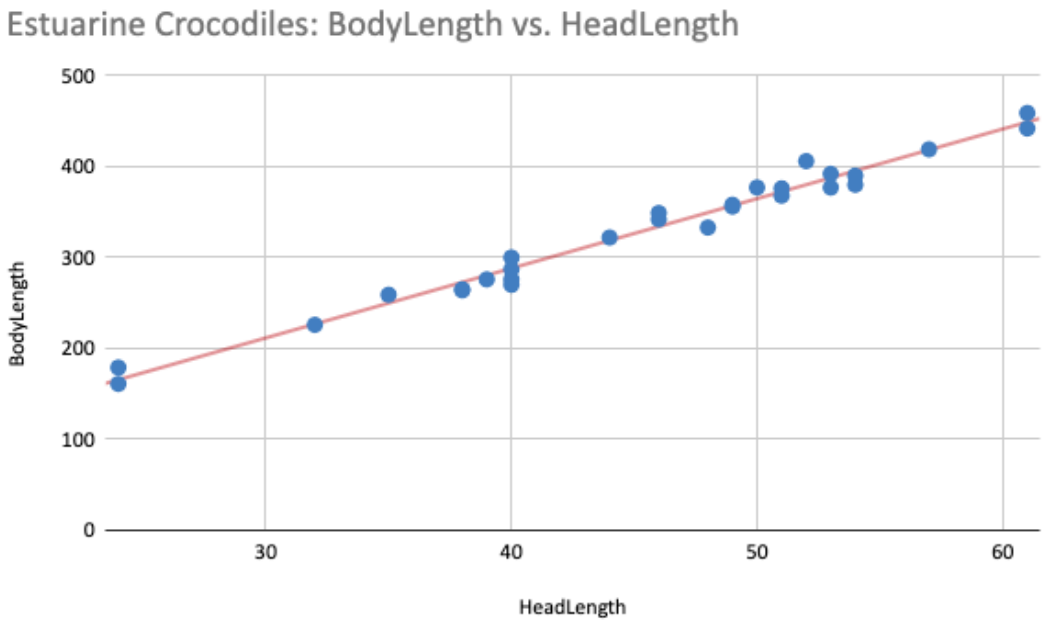
Week 11 - Linear Regression

Estuarine Crocodiles

- Data on estuarine (saltwater crocodiles)
- Sample size: 28 crocodiles.
- Variables: head length and body length.

We will try to predict the body length of the crocodile if we have only measured the length of its head.

Relationship between Body Length and Head Length

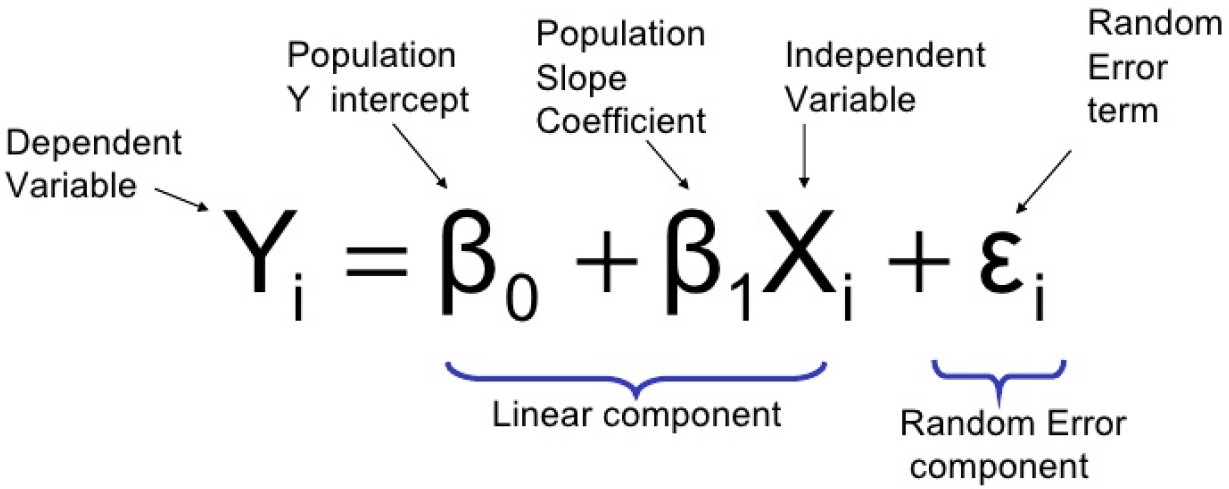


Linear Equations

$$Y = mX + b$$

where m is the slope and b is the intercept.

Simple Linear Regression

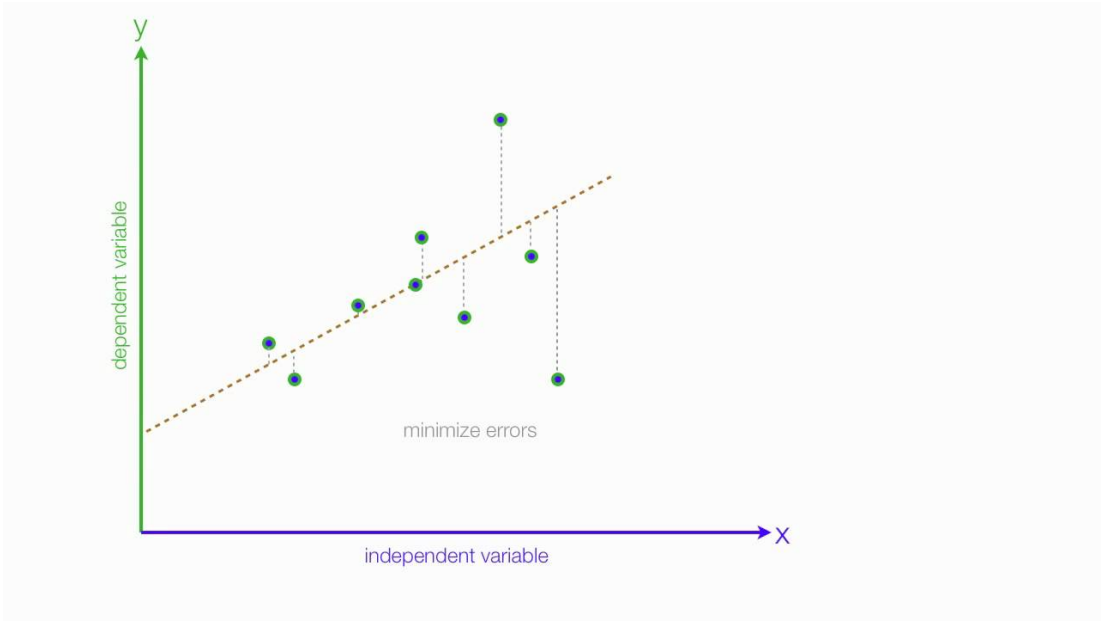


Simple Linear Regression

Find parameter values (also called **coefficients**, values for β_0 and β_1) such that the line

$$Y = \beta_0 + \beta_1 X + \epsilon$$

is as close to the data as possible.



OLS Assumptions

1. The conditional distribution of our error term, that is the distribution of our error term after we run the regression using X, needs to have an average of zero. This assumption is also referred to as *exogeneity* or *exogenous* regressors: our regressors cannot be correlated with the error terms.
2. In many cases, we will go one step further and assume that our error terms (after conditioning on regressors) are *normally distributed* with mean 0.
3. Our error are iid, which stands for independent and identically distributed. The idea behind this is that the errors we are making are random, and do not have a systematic correlation across different observations.
4. Large outliers in our variables are rare.

Predictions

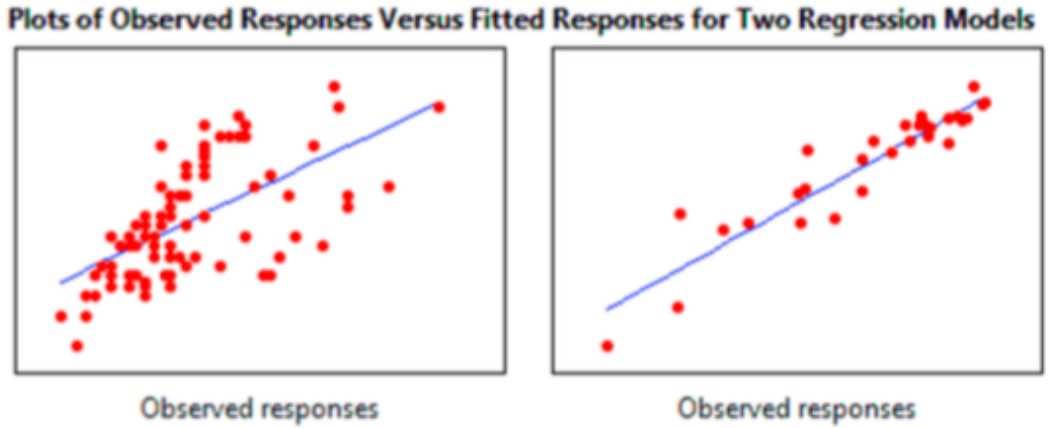
Once we have established the coefficient for our regression line, in a next step, we can use it to make a prediction for our variable of interest. In order to predict values for Y, we will use the equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Note that we do not include an error term in our prediction! We use the $\hat{Y}$ symbol to indicate that this is a *predicted* value for Y.

Goodness of Fit

We can say that a model fits the data well if the differences between the predicted values and observed values are small. Another way of saying this is that our model provides a good fit, if a large part of the variation in Y can be captured by our model, and only a small part remains unexplained.



The value for R^2 will always range between 0 and 1!

Standard Errors

- Assume that there is one true relationship between our variables, which we could find when perfectly knowing and having all of the data and the way it is related. Can we find it having only a small sample?

- Each of the coefficients will get also a standard error, which tells us how close it is to the true coefficient value. The standard error is a measure of **precision** on much our point estimate for the coefficient varies from its true population value.

- The larger the standard error, the less precise is the measurement of our coefficient.

- Closely related to the standard deviation, but adjusted by the size of our sample: the larger the sample (the closer the sample size to the population size), the smaller the standard error, eventually approaching zero as our sample size approaches full population size.

Confidence Intervals

- Coefficient is only a point estimate for the true population value, but measured with an error.
- Interval of values for which we're pretty sure that it contains the true population value: **confidence interval**
- Most commonly used confidence level: 95 %.

Confidence Interval

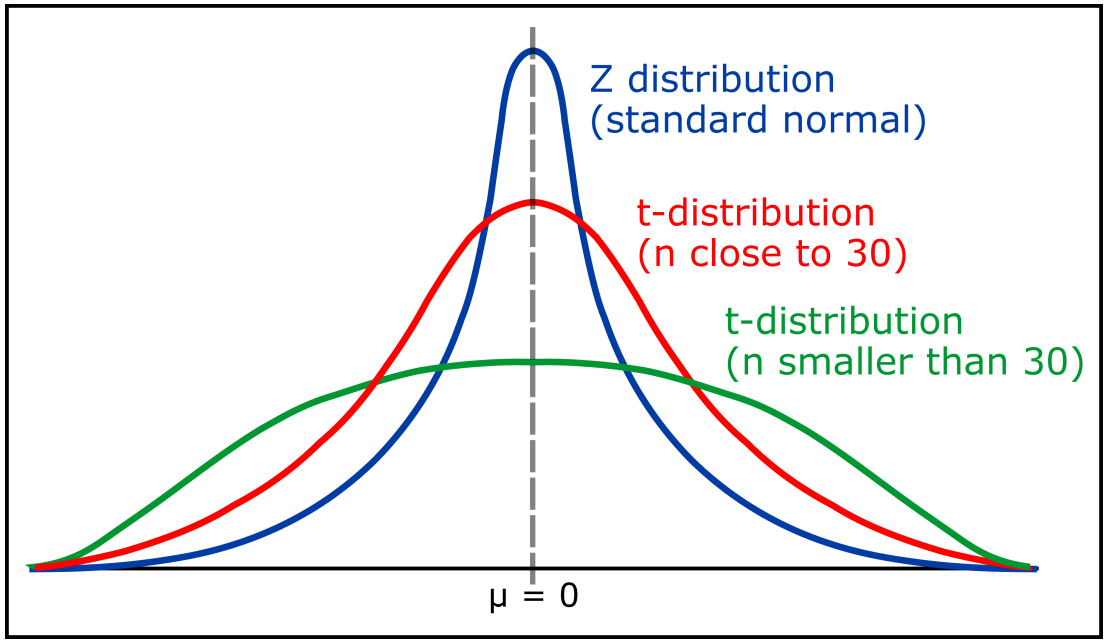
$$CI_{\beta_1} = \beta_1 + -(t - score) * StdError_{\beta_1}$$

where the t-statistic is taken from a student-t distribution with n - k degrees of freedom.

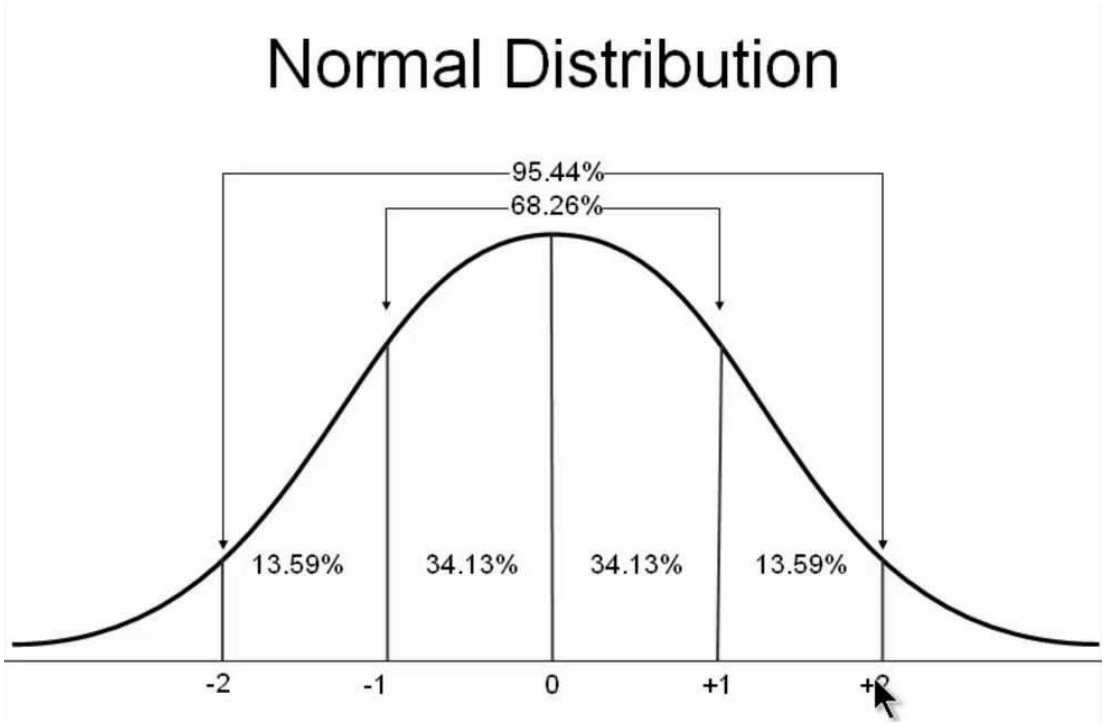
If our sample (and thus n) is large enough, the student-t distribution approaches a normal distribution, and we can use

$$CI_{\beta_1,0.95} = \beta_1 + -1.96 * StdError_{\beta_1}$$

Student T Distribution



Normal Distribution



Significance of Coefficients

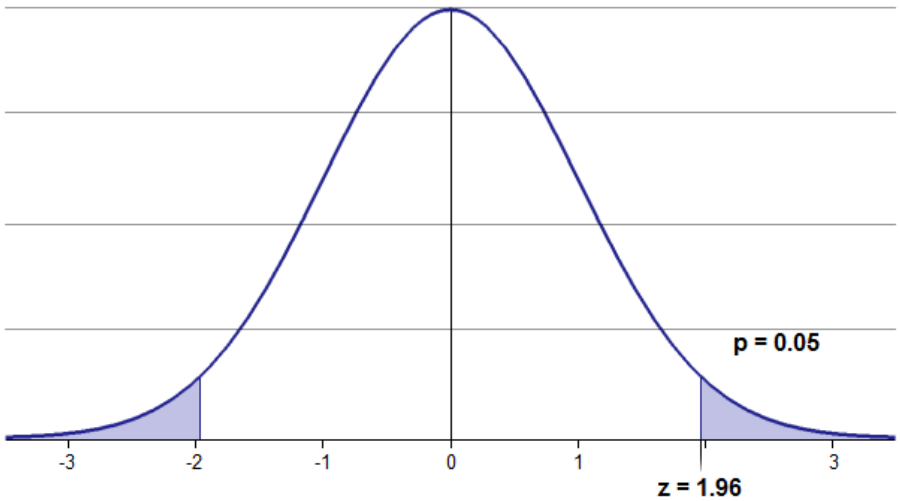
- Is the relationship between our variables is actually significant, and also likely to exist in the whole population in general?
- In other words: is the value of our coefficient in reality zero, that is the relationship does not exist?
- To determine this: is the number 0 is contained in our confidence interval?
- Second method: **hypothesis testing**.

Hypothesis Testing

- We formulate a specific hypothesis, and then see whether we have enough evidence to reject it. Very important: we never truly *accept* a hypothesis, we either *reject* it or fail to have enough evidence to reject it.

- If we want to test whether or not out coefficient is significantly different from zero, our null hypothesis will be: that it is equal to zero

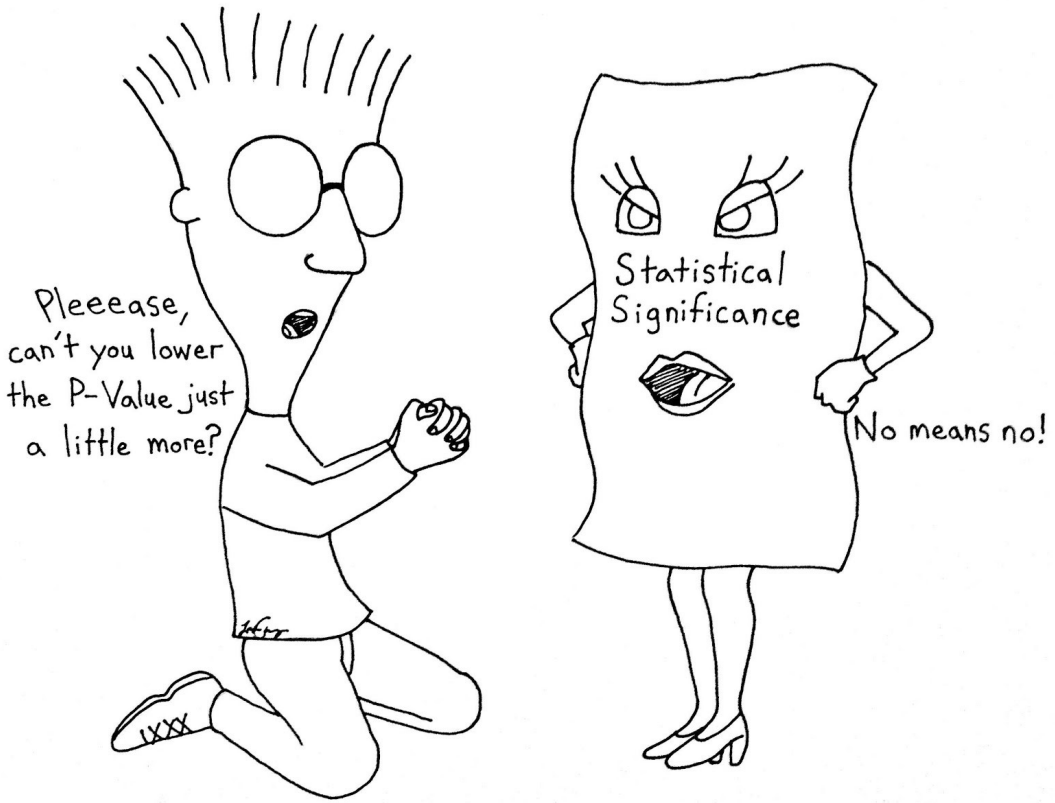
$$H_0 : \beta_1 = 0$$



Hypothesis Testing

- If the null hypothesis were true, what would be the probability to observe a value as large as the one we received purely by chance?
- If this probability is significantly small, then we have evidence to reject this null hypothesis.
- Create a threshold for how small we want this probability to be, and in line with our desired 95 % confidence interval commonly choose that we want this value to be smaller than 5%.

P-Values


Pleeease, can't you lower the P-Value just a little more?
Statistical Significance
No means no!

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \epsilon$$

where

- Y is the dependent variable
- k is the number of regressors, such that there are a total of k independent variables $X_1, ..., X_k$
- β_0 is the intercept
- β_1 is the coefficient for the regressor X_1, β_2 is the coefficient for the regressor X_2 etc
- ϵ is the regression error

Multiple Linear Regression

In a multiple regression, the coefficients will now represent *marginal* effects: β_1 tells us the effect on Y if we change X_1 by one unit, while keeping all the other X variables constant (ceteris paribus)! Similarly to before, in order to predict values for Y, we will use the equation

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_K X_K$$

OLS Assumptions

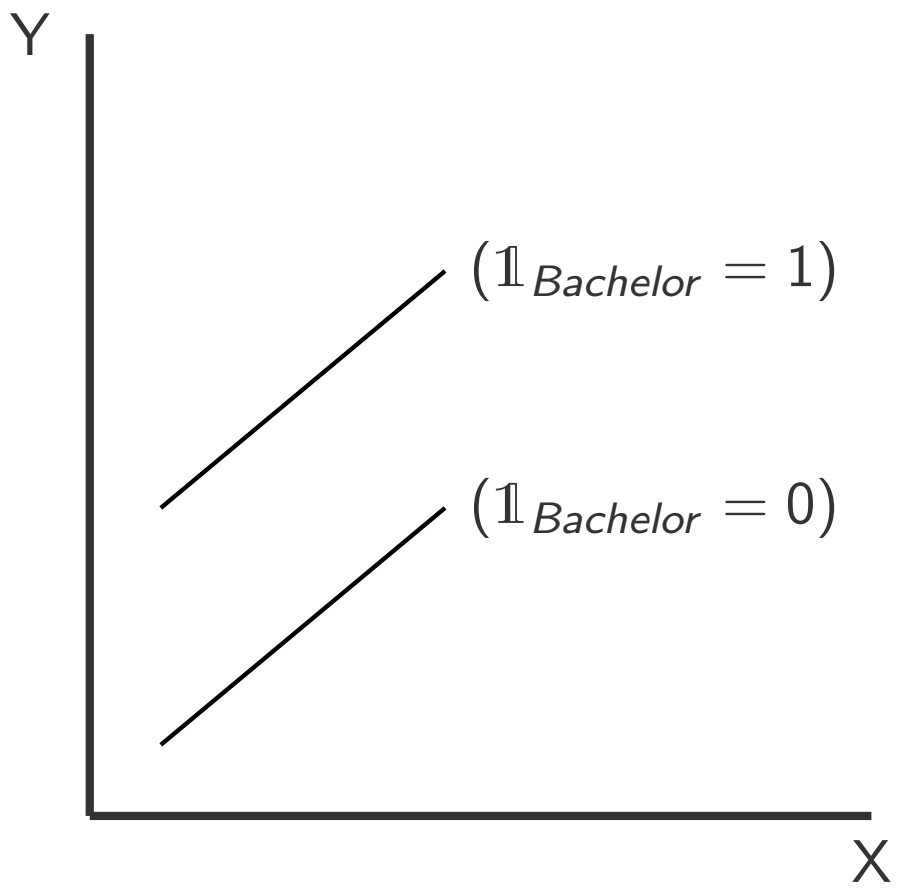
1. The conditional distribution of our error term, that is the distribution of our error term after we run the regression using X, needs to have an average of zero. This assumption is also referred to as *exogeneity* or *exogenous* regressors: our regressors cannot be correlated with the error terms.
2. In many cases, we will go one step further and assume that our error terms (after conditioning on regressors) are *normally distributed* with mean 0 and a constant variance.
3. Our error are iid, which stands for independent and identically distributed. The idea behind this is that the errors we are making are random, and do not have a systematic correlation across different observations.
4. Large outliers in our variables are rare.
5. There is no perfect multicollinearity among regressors.

Interpretation: Continuous Variables

For continuous (or discrete) variables, we can interpret our coefficient simply as "if I change X by one unit, by how many units does Y change"?

Interpretation: Binary Variables

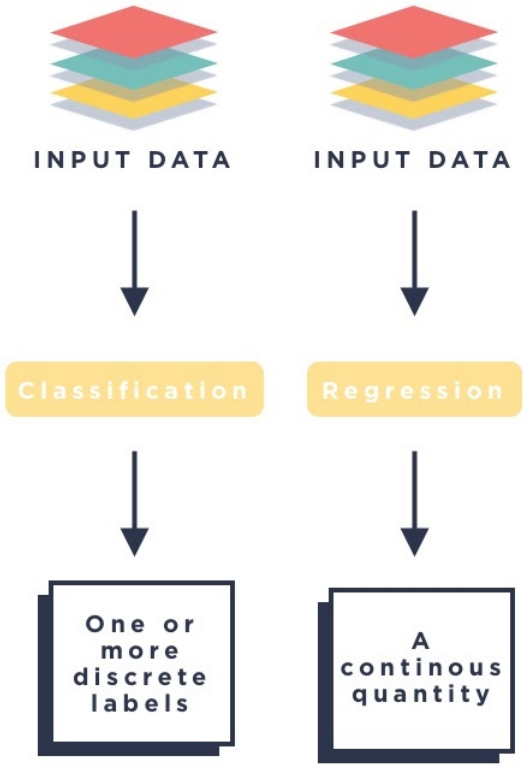
$$\text{HourlyEarnings} = \beta_0 + \beta_1 \text{WorkExperience} + \mathbb{1}_{Bachelor} + \epsilon$$



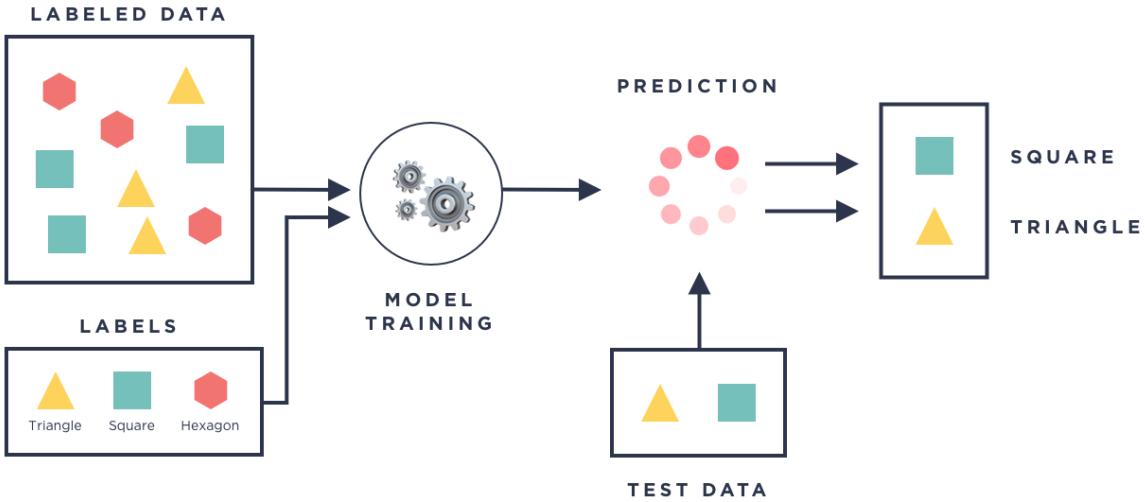
Interpretation: Categorical Variables

- Qualitative variables with several (more than two) categories.
- Example: highest educational achievement, i.e. with possible values high school, bachelor's degree, master's degree, PhD.
- Define a baseline category (similar to the dummy variable equal to zero) and then define dummy variables for the rest of categories! $\mathbb{1}_{bachelor}$, $\mathbb{1}_{master}$ and $\mathbb{1}_{phd}$
- Why can't we not include all four categories as variables? Multicollinearity!

Classification vs Regression



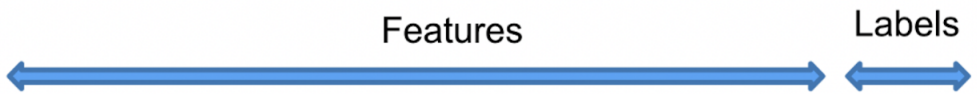
Classification Model



Regression Problem

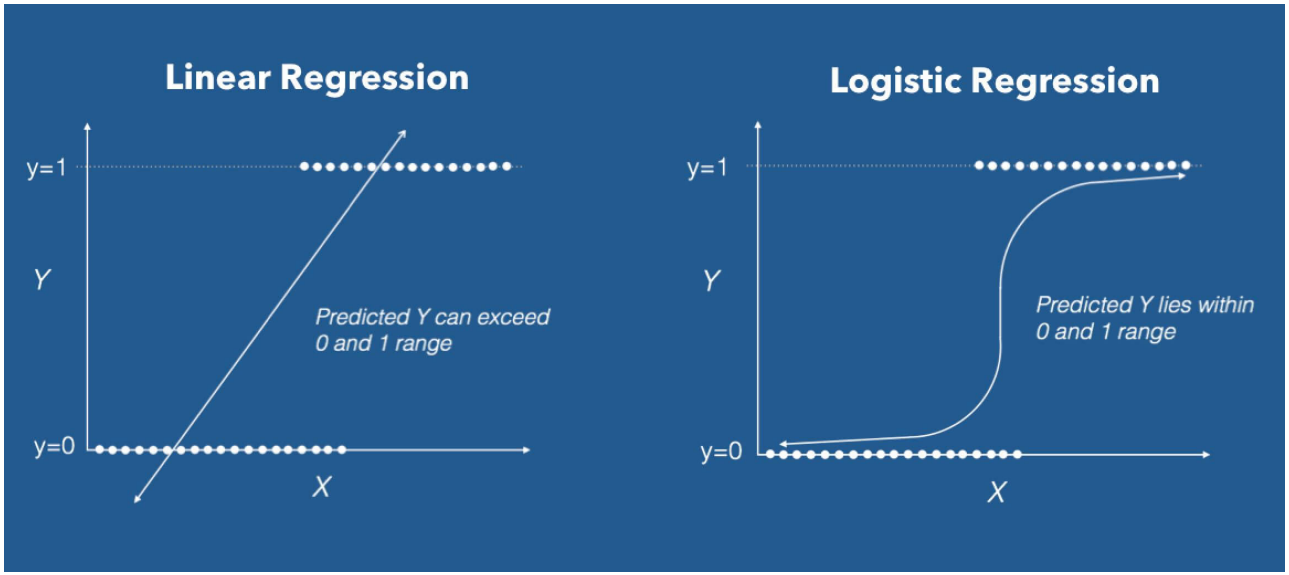
HV1	IC1	IC2	IC3	IC4	IC5	AVGGIFT	TARGET_D
2346	420	446	468	503	14552	15.5	21
497	350	364	357	384	11696	3.08	3
1229	469	502	507	544	17313	7.5	20
325	148	181	171	209	6334	6.7	5
768	174	201	220	249	7802	8.78571429	10
557	211	188	221	205	5550	13	16
2145	474	492	522	554	18340	11.5714286	15
2184	351	376	394	419	16480	12.5	20
1442	369	394	445	488	26462	7.84615385	10
1708	437	586	551	684	29098	9.76923077	20
1054	584	644	652	726	26074	13.5384615	20
1062	486	550	555	584	17908	15.3333333	20
849	457	508	470	519	16386	12.8	25
213	222	273	283	329	12227	5.125	5
574	289	318	315	363	11250	3.55555556	4
2506	449	455	501	517	16302	8.875	50
622	347	378	401	416	15808	15	25
764	272	361	346	424	16257	7.91304348	15
681	335	398	356	419	14011	30.75	51

Classification Problem



	loan_id	account_id	date	amount	duration	payments	status
0	5314	1787	930705	96396	12	8033.0	B
1	5316	1801	930711	165960	36	4610.0	A
2	6863	9188	930728	127080	60	2118.0	A
3	5325	1843	930803	105804	36	2939.0	A
4	7240	11013	930906	274740	60	4579.0	A
5	6687	8261	930913	87840	24	3660.0	A
6	7284	11265	930915	52788	12	4399.0	A
7	6111	5428	930924	174744	24	7281.0	B
8	7235	10973	931013	154416	48	3217.0	A
9	5997	4894	931104	117024	24	4876.0	A
10	7121	10364	931110	21924	36	609.0	A
11	6077	5270	931122	79608	24	3317.0	A
12	6228	6034	931201	464520	60	7742.0	B
13	6356	6701	931208	95400	36	2650.0	A
14	5523	2705	931208	93888	36	2608.0	A
15	6456	7123	931209	47016	12	3918.0	A

Logistic Regression



Logistic Regression