

CHEAT SHEET ON REGRESSIONS (in progress)

1 Simple Linear Regression

1.1 Univariate Model (One explanatory variable only)

$$y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$$

The Ordinary Least Squares approach is designed to minimize the magnitude of the estimated residuals ($\sum_{i=1}^n \epsilon_i^2$). This method, in the univariate case gives the following estimates of β_0 and β_1 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{Y})(x_i - \bar{X})}{\sum_{i=1}^n (x_i - \bar{X})^2} = \frac{n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} \text{ and } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

1.2 Including Logarithms

$$\ln y_i = \beta_0 + \beta_1 * x_i + \epsilon_i$$

Here, β_1 is the proportionate change in y arising from a unit change in x (log-linear model).

$$y_i = \beta_0 + \beta_1 * \ln x_i + \epsilon_i$$

Here, β_1 is the change in unit in y arising from a proportionate change in x (linear-log model).

1.3 Multivariate Model (2 or more explanatory variables)

$$y_i = \beta_0 + \beta_1 * x_{1,i} + \beta_2 * x_{2,i} + \dots + \beta_k * x_{k,i} + \epsilon_i$$

In this case, the formulas for the $\hat{\beta}_i$ are much more complex. You do not need to know them by heart!

1.4 Hypothesis Testing on a Single Coefficient

1.4.1 One-sided test

$$\begin{aligned} H_0 : \beta_1 &= \alpha \\ H_1 : \beta_1 &\neq \alpha \end{aligned}$$

We will reject the null if $\hat{\beta}_1$ is "far" from α . So we construct the Z-statistic:

$$Z = \left| \frac{\hat{\beta}_1 - \alpha}{\sigma_{\hat{\beta}_1}} \right| \text{ and compare it to } Z_{1-\alpha/2}$$

and we reject the null if the Z-statistic is greater than the threshold we are interested in (let's say for the significance level of 5%, $Z > 1.96$).

1.4.2 Two-sided test

$$\begin{aligned} H_0 : \beta_1 &= \alpha \\ H_1 : \beta_1 &> \alpha \end{aligned}$$

Here, we construct the Z-statistic:

$$Z = \frac{\hat{\beta}_1 - \alpha}{\sigma_{\hat{\beta}_1}} \text{ and compare it to } Z_{1-\alpha}$$

and we reject the null if the Z-statistic is greater than the threshold we are interested in (let's say for the significance level of 5%, $Z > 1.65$).

1.4.3 Student (t-) Distribution

Alternatively, we could use the student t-table if we have a small population ($n < 120$), and construct the following t-statistic:

$$t = \frac{\hat{\beta}_1 - \alpha}{\sigma_{\hat{\beta}_1}}$$

and we reject the null if the t-statistic is greater than the threshold we are interested in ($t_{n-(k+1), 1-\alpha/2}$ in the two-sided case, and $t_{n-(k+1), 1-\alpha}$ in the one-sided case).

2 Hypothesis Testing on Several Regression Coefficients: F-Test

We form the original, *unrestricted* regression.

$$y_i = \beta_0 + \beta_1 * x_{1,i} + \beta_2 * x_{2,i} + \dots + \beta_k * x_{k,i} + \epsilon_i$$

Now we want to test the following hypothesis:

$$H_0 : \beta_1 = \alpha \text{ and } \beta_2 = \gamma$$

$$H_1 : \text{not } H_0$$

We form an extra regression, called the *restricted*-regression:

$$y_i - \alpha * x_{1,i} - \gamma * x_{2,i} (\equiv y_i^*) = \beta_0 + \beta_3 * x_{3,i} + \dots + \beta_k * x_{k,i} + \epsilon_i$$

and we compare the R^2 of both restricted (res) and unrestricted (unres) regressions. Denoting m the number of restrictions, we form the following ratio:

$$F = \frac{(R_{unres}^2 - R_{res}^2)/m}{(1 - R_{unres}^2)/(n - (k+1))}$$
 and compare with $F_{m, n-(k+1)}$

We compare that ratio to the critical value that we are interested in, and if the ratio is greater than this threshold, we reject the null hypothesis.

3 Dummy Variables

Dummy variables allow the intercept of the regression to vary for different groups in the population. For instance we could create the following dummy variable:

$$\text{gender}_i = \begin{cases} 1 & \text{if individual } i \text{ is a female} \\ 0 & \text{if individual } i \text{ is a male} \end{cases}$$

and run the regression:

$$\text{earnings}_i = \beta_0 + \beta_1 * \text{experience}_i + \beta_2 * \text{gender}_i + \epsilon_i$$

where β_0 is the intercept for the male group, $\beta_0 + \beta_2$ is the intercept for the female group and β_1 is the slope for both regression lines. Here only the intercept changes, the slope remains the same across groups (so we have parallel regression lines).

We can also create dummy variables for populations divided in more than one group. For instance, if the population is divided in 5 educational groups (they *need* to be mutually exclusive), we can create 5-1=4 dummies indicating 4 out of the 5 educational groups and include them in the regression:

$$\text{earnings}_i = \beta_0 + \beta_1 * \text{experience}_i + \beta_2 * \text{out}_i + \beta_3 * \text{hschool}_i + \beta_4 * \text{somec}_i + \beta_5 * \text{college}_i + \epsilon_i$$

the reference group being *graduate*. Here β_0 is the intercept of the reference group, in our case, the graduate group. $\beta_0 + \beta_2$ is the intercept for school drop-outs, $\beta_0 + \beta_3$ is the intercept for high-school graduates, $\beta_0 + \beta_4$ is the intercept for students who went to college without finishing and $\beta_0 + \beta_5$ is the intercept for college graduates. Once again all the regression lines are parallel.

4 Interaction Variables

Dummy variables can be interacted with other variables to change also the slope across groups. For instance we could create the following interaction:

$$earnings_i = \beta_0 + \beta_1 * experience_i + \beta_2 * gender_i + \beta_3 * experience_i * gender_i + \epsilon_i$$

where β_0 is the intercept for the male group, $\beta_0 + \beta_2$ is the intercept for the female group and β_1 is the slope for the male group, and $\beta_1 + \beta_3$ is the slope for the female group. Here not only does the intercept changes, but the slopes also (we don't have parallel regression lines anymore).

5 Multicollinearity

5.1 Including an Irrelevant Variable

Basic idea: it DOES NOT change anything to the estimated values of the other β s...

5.2 Omitted Variable Bias

On the other hand, if you forget to include a variable that is important in your regression, your estimated β s will be biased. Imagine that you are trying to explain weight at birth with the pregnant mother's health status. You run the following regression:

$$weight_i = \beta_0 + \beta_1 * smoke_i + \epsilon_i$$

Unfortunately, you have forgot to include the consumption of alcohol of the mother, that is highly correlated with the mother's smoking habit ($alcohol_i = \gamma * smoke_i + \nu_i$). The true model would be:

$$weight_i = \delta_0 + \delta_1 * smoke_i + \delta_2 * alcohol_i + \epsilon_i$$

and your β_1 is estimating not δ_1 but $\delta_1 + \delta_2\gamma$. So in your estimate of the impact of smoking on weight at birth, you are including the impact of alcohol consumption as well...

5.3 Nonlinearity

So far, we have always assumed that the relationship between the explanatory variables and the outcome variable was linear (a straight line). But it may be the case that in some regressions, we do not want to model a linear relationship (the returns to education for instance on earnings). The easiest way to introduce such a non-linear relationship is to introduce polynomials on the RHS (right-hand side):

$$\ln y_i = \beta_0 + \beta_1 * x_i + \beta_2 * x_i^2 + \epsilon_i$$

where β_2 is the rate of change of the slope, while the slope is a function of both β_1 and β_2 . Thus, the slope may increase (decrease) at an increasing (decreasing) rate.

We could also use logarithms (as the relationship between a variable and its natural logarithm is a non-linear one).

5.4 Multicollinearity

One of the fundamental assumptions of the multiple regression model is that the x's are not collinear. What happens if you include collinear variables? Imagine for instance that every time the weather is nice, (a) you go to the pool and (b) you don't take the bus to school but you walk. Now, you usually get more tired those evenings. Because both (a) and (b) always happen together, it is impossible to disentangle the effect of swimming from the effect of walking on your tiredness. Just because both events are perfectly correlated. So you can't estimate separately the effect of swimming and that of walking.

You can also have multicollinearity if one variable takes the same value for the whole population (all pregnant persons are women), or if you include the whole set of dummy variables (instead of (k-1)).