**What do we look at in regression output?**

1. First, we look at the F-statistic and the associated p-value. The F-test is the so-called *omnibus* test, which tests the null hypothesis that none of the independent variables in the model is a significant predictor of the dependent variable, against the alternative hypothesis that *at least* one of the independent variables is a significant predictor of the dependent variable. A model where we cannot reject the null hypothesis is generally a bad model.
   1. As an aside, it *is* possible, in a multiple regression model, for one or more of the predictors to be significant, but for the model F-statistic to not be significant. This happens fairly rarely, but in that instance, the test does not affect conclusions that may be drawn from the variables that are significant. For more information, see <http://en.wikipedia.org/wiki/Omnibus_test>.
   2. In a *simple regression*, it will always be the case that if the predictor is significant, so is the F-statistic, and if the predictor is not significant, neither is the F-statistic.
2. Next, look at the p-value associated with each independent variable. If p-value for a certain independent variable is less than 0.05, we can reject the null hypothesis that this particular predictor is *not* a significant predictor of the dependent variable.
3. If the predictor is significant, we look at the sign and value of the for that predictor. The sign tells us whether the relationship between the predictor and the dependent variable is positive or negative (i.e., direct or inverse), and the value of tells us the amount by which the dependent variable changes when the value of the predictor increases by 1 unit, holding the values of other predictors constant\*.
4. We also look at R2, which is calculated as 1 – SSE/SST, where the SSE is the sum of squared residuals and SST is the total variability in the dependent variable. R2 is called the *coefficient of determination* and higher values are indicative of a better model, as the value of R2 is the proportion of variance in the dependent variable that has been explained by the model.
5. Finally, we look at the adjusted R2, which is the R2 adjusted for the number of predictors in the model. Again, higher values of adjusted R2 are better.

\* What do we mean by “holding other predictors constant”?

Imagine we regress our dependent variable, ***MEDHHINC***, on three predictors (***MEDHVAL***, ***PCTVACANT*** and ***PCTBACHMOR***) and that the coefficient of ***MEDHVAL*** is estimated as (or $500). That is, for every 1 unit (i.e., $1,000) increase in ***MEDHVAL***, the dependent variable ***MEDHHINC*** will go up by , or $500, controlling for other predictors (**PCTVACANT** and **PCTBACHMOR**) \*\*, or holding these other predictors (**PCTVACANT** and **PCTBACHMOR**) constant.

So what does *holding constant* mean?

Imagine we have 2 block groups, A and B, with the following data:

|  |  |  |
| --- | --- | --- |
|  | Block Group A | Block Group B |
| **PCTVACANT** | 6% | 6% |
| **PCTBACHMOR** | 71% | 71% |
| **MEDHVAL** | $150k ($150,000) | $151k ($151,000) |

The two block groups A and B are identical on the variables ***PCTVACANT*** and ***PCTBACHMOR*** and differ on ***MEDHVAL*** (by $1k, or $1000). So, compared with block group A, the dependent variable ***MEDHHINC*** in block group B is estimated to be higher by , or $500, because ***MEDHVAL*** is higher by 1 unit ($1000), and the values of the other predictors remain the same (i.e., are held constant).

Now, let’s work with an example where the dependent variable and predictor are log-transformed. We may interpret our beta coefficients in terms of percentages, as described in detail elsewhere.

Imagine that we now regress our dependent variable, ***LNMEDHHINC***, on 3 predictors (***LNMEDHVAL***, ***PCTVACANT*** and ***PCTBACHMOR***) and that the coefficient of ***LNMEDHVAL*** is estimated as . That is, for every 1% increase in ***MEDHVAL***, ***MEDHHINC*** goes up by 0.5%, holding the other predictor predictors constant.

Imagine we have 2 block groups, A and B with the following data:

|  |  |  |
| --- | --- | --- |
|  | Block Group A | Block Group B |
| **PCTVACANT** | 6% | 6% |
| **PCTBACHMOR** | 71% | 71% |
| **MEDHVAL** | $150k ($150,000) | $150k + 1% of $150k ($151,500) |

Imagine further we have 2 other block groups, C and D with the following data:

|  |  |  |
| --- | --- | --- |
|  | Block Group C | Block Group D |
| **PCTVACANT** | 6% | 6% |
| **PCTBACHMOR** | 71% | 71% |
| **MEDHVAL** | $100k ($100,000) | $100k + 1% of $100k ($101,000) |

The two block groups A and B are identical on the variables ***PCTVACANT*** and ***PCTBACHMOR*** and differ on ***MEDHVAL*** by 1%. So, compared with block group A, the dependent variable ***MEDHHINC*** in block group B is estimated to be higher by %, or 0.5%. The same thing holds when comparing block groups C and D – they are identical on the variables ***PCTVACANT*** and ***PCTBACHMOR*** and differ on ***MEDHVAL*** by 1%. So, compared with block group C, the dependent variable ***MEDHHINC*** in block group D is estimated to be higher by %, or 0.5%. (Because we’re employing logarithms, a 1% increase will be different depending on the initial value when we translate it to $ amounts. When the initial value is $100,000, a 1% increase will correspond to $1,000. When the initial value is $150,000, a 1% increase will correspond to $1,500. This is one of the caveats of a non-linear transformation.) But we can say that – regardless of the initial value of the predictor – a 1% increase in it will, on average, result in a .5% increase in the value of the dependent variable, when the values of other predictors are unchanged. We can also say that, the value of , which is our sample estimate of the population parameter , is the average expected (percentage) change in the dependent variable when the predictor goes up by 1%.

\*\* Controlling vs. Holding constant: “In statistics, **controlling** for a **variable** is the attempt to reduce the effect of confounding **variables** on an observational study. It means that when looking at the effect of one **variable**, all other **variable** predictors are held constant.” (Source: <https://en.wikipedia.org/wiki/Controlling_for_a_variable#cite_note-1>)

# Assumption

1. Linear relationship between dependent variable *y* and each of the predictors *x*.
2. Normality of residuals. This is an assumption that isn’t as important as some of the others, especially if we have a large sample size, but most statisticians still like to make sure that it is met whenever possible.
3. Homoscedasticity – the variance of the residuals is constant regardless of the values of each *x* (or the values of , i.e., values of *y* predicted by the model).
4. Independence of observations. That is, there should be no spatial, temporal or other forms of dependence in the data. In this course, we will focus specifically on a lack of spatial dependencies.
5. No multicollinearity (this applies only to multiple regression). That is, predictor variables shouldn’t be strongly correlated with each other.

6. No fewer than 10 observations per predictor

**Assumption Checks**

1. Scatterplots between *y* and each of the predictors *x*.
2. Look at the histogram of residuals to see if they are normal. As an aside, we are generally more likely to see normal residuals if the dependent variable is normal and the predictor variables are normal (or binary).
3. Look at scatterplots of standardized residuals against each predictor and see if variance of residuals remains the same for different values of each predictor. When there is more than one predictor, it’s simpler to plot standardized residuals against values of , and see if variance of residuals remains the same for different values of . If not, we have *heteroscedasticity*, which we should address, if possible.
4. Look at the Moran’s I of the residuals, or the values of *y* to examine whether the regression residuals, or the dependent variable itself, are spatially autocorrelated.
5. Look at correlation matrix of predictors. If correlation between any pair of predictors > 0.8 (or < - 0.8), we might have multicollinearity. Alternatively, regress each predictor on remaining predictors and look at each of the R2 (shouldn’t be more than 0.8) or each of the Variance Inflation Factors (shouldn’t be more than 4).

6. Look at ratio between # of observations and

# of predictors

**Possible Fixes If Assumption Not Met**

1. Transform variables (e.g., log); run non-linear (e.g., polynomial) model.
2. Remove outliers; transform variables (e.g., log-transformation).
3. Presence of heteroscedasticity often means that there is systematic under-or over- prediction happening in the model (we want this to not happen, and for any errors to be *random*). The inclusion of additional predictors, running a spatial regression (and sometimes variable transformations and removal of outliers) may help reduce or eliminate *heteroscedasticity* (again, we want *homoscedasticity*).
4. If Moran’s I of the residuals or dependent variable is significant, run a spatial regression (e.g., spatial lag, spatial error, geographically weighted regression) instead of OLS regression.
5. If 2+ predictors are strongly correlated, include only one of them in the regression; including all of them will lead to incorrect parameter estimation. Alternatively, run ridge or lasso regression.

6. Remove some predictors. Alternatively, run

ridge or lasso regression.