1 PLS 300: Optional notes on regression analysis

1.1 Regression Estimation

We will not get too far into the technical details of estimating the multiple regression model, beyond our conceptual understanding of the OLS regression plane (in 3D) minimizing the sum of squared residuals, \((y_i - \hat{y}_i)^2\). Those of you with an interest in learning more about estimation of the regression parameters should consult an additional textbook, as the textbook does not go into any detail on the subject. But the main point is thus that we are left with a fitted least squares regression equation:

\[ \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k \]

2 PLS 300 KILBURN

2.1 Notes: The Regression Model and The Error Term

As a model of the population, we can never capture the dependent variable entirely as a function \(y = f(X, \beta)\), thus we incorporate an error term, \(\epsilon_i\), into the equations of section 2.1. The error represents non-systematic influence on the dependent variable, which could be due to 3 sources: 1) omitted variables; 2) idiosyncratic variation in \(y_i\); and 3) measurement error in \(y_i\). Through the error term we acknowledge that while there are other influences on our dependent variable, they are not important enough to emphasize in the model. The inferences that we are able to draw from the regression model depend in part upon the assumptions we are willing to make about the behavior of these errors.

2.1.1 Critical Assumptions: the Model, Error, Predictors, and Observations

Our ability to draw inferences from the regression depend upon the extent to which we can assume that certain conditions (assumptions) are met:

**Assumption 1:** The regression model relating the dependent to independent variables is assumed to be linear in the regression parameters. We can easily check this linearity assumption by examining scatter plots of \(Y\) versus \(X\) predictor variables. Since multiple regression models possess a high dimensionality, this isn’t always practical. (Other techniques are discussed below).

**Assumption 2:** Each independent, predictor variable \(X_i\) is fixed and non-random. Strictly speaking, this assumption is violated if any independent variables possess measurement error, or if the researcher lacks control over values of \(X_i\) as in the case of experimental controls. Nonetheless, for our purposes this is not a critical assumption. We are not interested in explaining variation in \(X_i\); we assume that our \(X_i\) is fixed and concentrate our efforts on explaining variation in \(Y\).

**Assumption 3:** All \(\epsilon_i\) are normally distributed and therefore continuous. While this assumption is not necessary for estimation of the OLS regression parameters, it is necessary for hypothesis testing.
Assumption 4: The errors have a mean of zero, or specifically, the expected value of $\epsilon_i$ given $X_i$ is zero; $E[\epsilon_i|X_i] = 0$. We adopt this assumption in order to assure ourselves that we have non-systematic errors; the non-systematic forces pushing up values of $Y_i$ are cancelled out by forces pushing $Y_i$ downward. Thus, if the errors cancel out, we can model our dependent variable $Y$ as a function of the $\beta_0 + \beta_k X_i$ etc.

Assumption 5: This assumption is critical: the errors $\epsilon_i$ and independent predictor variables are uncorrelated. In other words — whatever we left out of the right-hand-side of the model as independent variables is not correlated with what we included. This assumption is essential, since if the errors and the predictors are correlated then one cannot estimate their individual effects on the dependent variable. Thus, we have to be confident that there is no missing or omitted variable bias.

Assumption 5: The errors are homoskedastic. That is, constant variance for $\epsilon_i$ across all levels of the predictor variables. Where this assumption is violated, we say the errors are ‘heteroskedastic’. Heteroskedasticity is usually a problem with cross-sectional data, especially clustered data collected across groups of individuals, nations, etc. The assumption of homoskedasticity has important implications for hypothesis tests on regression parameters, discussed further below. Formally, we represent this assumption as $\text{Var}(\epsilon_i) = \sigma^2$

Assumption 6: There is no autocorrelation present among the errors, formally $E(\epsilon_i \epsilon_j) = 0$ for $i \neq j$. We assume that the correlation between disturbances should be zero. In other words, observations on the dependent variable are independent of one another, where knowing the value of one observation does not help in predicting another. —This assumption is often violated with time series data; presidential approval at time $t$ is related to approval at $t+1$, etc. This assumption, too affects hypothesis testing.

Assumption 7: All observations have an approximately equal role in influencing the regression. Violations of this assumption may occur when data are grouped across clusters, necessitating a correction for heteroskedasticity.

These assumptions are a critical part of the regression model for the population. How do we know about the population? —By studying the sample regression model. As long as these assumptions reasonably hold, we are assured that our estimated population regression line holds reasonably well. In a nutshell, as long as these assumptions hold reasonably well, then OLS is BLUE. Or in other words, our Ordinary Least Squares Estimator for multiple regression is the Best Linear Unbiased Estimator of our population parameters. If these assumptions are tenuous at best or fail, then our regression model is Blah, not BLUE.

### 2.2 General Diagnostics

Sometimes our assumptions about the error term may not be tenable:

**Heteroskedasticity – Violating Assumption 5** A principle problem of heteroskedasticity is deflated standard errors, making it appear that predictors are significant when they are in fact, not. A visual method of assessing heteroskedasticity is by visually examining the distribution of residuals. When the residuals at each level of the predictor variables have unequal variation, this is heteroskedasticity. When we plot standardized residuals
on the y-axis, and standardized predicted values on the x-axis, we want to observe ‘white noise’ residuals with no clear pattern; deviations from no pattern are cause for concern. Note that heteroskedasticity is almost always a problem with cross-sectional data.

NOTE: Heteroskedasticity is usually a problem with cross-sectional data that are ‘clustered’ – with cases across nations, regions, etc. or people within groups. Including dummy variables for these groups may help with the problem.

Autocorrelation – Violating Assumption 6  Autocorrelation means that between observations, errors are correlated. We can check for the presence of autocorrelation by examining whether residual errors across adjacent observations are correlated.

Multicollinearity – A problem of Too Little Information  Collinearity between two variables occurs when one predictor is too closely related to another, thus it becomes difficult to isolate the effects of each. While it is possible to have collinearity between variables that are moderately correlated, a good first check is to examine a correlation matrix for any variables that are correlated at least .90. A next step in diagnosing multicollinearity is to observe whether standard errors of beta coefficients change as predictors are entered into or taken out of a model, since a principle problem of collinearity is inflated standard errors.

2.3 Hypothesis testing and Model Comparison

Without hypothesis testing, multiple regression analysis is not particularly useful for empirical testing of theories. There are two possible tests: 1) tests of individual model parameters; and 2) tests of multiple parameters, also known as ‘block’ F-tests.

Tests of individual parameters — These are tests similar to the case of a bivariate regression model. We are interested in tests for a particular predictor in which we formulate a hypothesis concerning its effect in the population and this hypothesis is evaluated against the sample data. But to do so, however, we must be willing to assume that all $\epsilon_i$ of the model are identically, and independently, normally distributed.

Most often, we are interested in the following hypotheses:

$$H_0 : \beta_j = 0$$

$$H_a : \beta_j \neq 0$$

Thus, we hypothesize that a particular predictor, $X_j$ has no effect on the dependent variable in the population. The relevant test statistic:

$$T = \frac{\hat{\beta}_j}{\text{estimated standard error of } \beta_j} t_{n-k-1}$$
tk —how to calculate the standard error for a beta coefficient

We assess the probability that the absolute value of $T$ is greater than the value of $T$ in the sample. To find this probability, (a p-value) we relate $T$ to a student’s t-distribution with $n-k-1$ degrees of freedom. As usual, if $t < \alpha$, then $H_0$ is rejected; otherwise we fail to reject the null hypothesis $H_0$.

2.4 Interpretation of Partial Regression Coefficients

Thus, interpretation of partial regression coefficients is similar to interpretation of simple regression coefficients. Thus our estimate of $\beta_0$, termed ‘Beta-hat-zero’ $\hat{\beta}_0$ provides the expected value of the dependent variable assuming that all of the predictors take on a value of 0. Our estimates of $\hat{\beta}_j$ for all $j$ greater than 0 give the effect of the $j$th predictor, holding all else constant. This is the expected change in $y$ for a one unit change in the $j$th predictor, assuming no change in other predictors. (Note that even if predictor variables are correlated, and we would normally expect one predictor to accompany a change with the other, the interpretation of partial regression coefficients requires us to make this assumption of no change.) Comparing an ‘effect size’ of partial regression coefficients is generally not possible without some forethought about how the variables are scaled or without a transformation of the regression coefficients. The magnitude or partial regression coefficients is determined mostly by the scale of the independent variables, and are thus in general, independent variables are not usually measured on the same scale. In order to make regression coefficients comparable we can pursue two strategies. Generally, the first strategy — standardization of the scaling of predictors — is better. I think most researchers agree that the second strategy should be avoided. But we review it here, as in Agresti and Finley just for understanding.

2.4.1 standardization of scaling of predictors

Generally, the most preferred alternative is to facilitate a comparability by placing the predictors on a 0-1 scale wherever it makes sense to do so. As we discussed in class, in some cases this makes little substantive sense. So you should only apply it when the predictor variable is set on an arbitrary scale, or one for which an interval or ordinal scale would has no inherent meaning.

When it does make sense to do so, one can simply use the following transformation:

$$x_{ij}^* = \frac{x_{ij} - \min(x_j)}{\text{range}(x_j)}$$

Again, the principle advantage of this strategy is that the interpretation of partial regression coefficients is intrinsically meaningful: a partial regression coefficient is the expected change in $y$ for a change in the the entire scale of the predictor, holding constant all other predictors. Researchers can further interpret the magnitude of coefficients in terms of widths of the scale, by dividing coefficients in half for the magnitude of a one-half scale change in the predictor variables, or dividing by four for a one-fourth scale change, etc.
2.4.2 standardization of regression coefficients

A second strategy, one you should apply with caution, is to standardize the regression coefficients, as described in Agresti and Finley Chapters 9 and 11. Our standardized regression coefficients are obtained through the following transformation:

\[ \hat{\beta}_j^s = \hat{\beta}_j \frac{S_{x_j}}{S_y} \]

Multiplication of \( \hat{\beta}_j \) with a function of the standard deviation of \( x_j \), the scale of \( x_j \) is taken out of the slope estimate.

Standardized regression coefficients are in some cases indistinguishable from correlations. From any bivariate regression:

\[ \hat{\beta}_j^s = \hat{\beta}_j \frac{S_{x_j}}{S_y} = \frac{s_{xy}}{s_x s_y} = r_{xy} \]

While it is easy to rely upon these standardized coefficients, you should be wary of applying them. The interpretation of the coefficients in standard deviation units is unintuitive, and can be misleading in subgroup analysis — the analysis of regression results from sub-samples of your data. This misleading results occurs because while subgroups may have the same regression coefficient they may also have very different variances on the dependent variable. So relying on the magnitude of the standardized coefficient may result in a distorted view of hat the relative size of each regression coefficient really is.

2.5 Using Categorical (Nominal) Predictors in Regression Analyses

The treatment and interpretation of categorical versus continuous (interval or ratio, and even ordinal) scale predictors is fundamentally different. While for continuous variables we can assess the expected change in \( y \) for a given unit change in \( x \), this statement requires some additional clarification for categorical predictors. For instance, we need to specify what the unit change is. And whether different changes in \( x \) units mean the same thing.

Thus we model effects of categorical predictors as shifts in intercepts. The key insight is to conceptualize categorical predictors as a series of dummy (dichotomous) variables. A clear example of this is when the categorical variable of interest is strictly dichotomous, such as gender. Let us consider an example similar to the on discussed in class, the estimated impact of gender and education on a person’s salary. Here, as in all cases, we want to make sure that the dichotomous variable is coded from 0 to 1, where 0 is referred to as the ‘baseline’ against which the other category is compared.

Thus we have: \( y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \epsilon_i \)

Here \( X \) is the continuous predictor and \( D \) is a dummy (categorical, dichotomous) predictor. Image that \( D=1 \) for women and \( D=0 \) for men. Then we write out specific equations for each:
**Men** (D=0):  
\[ y_i = \beta_0 + \beta_1 x_i + \beta_2 * 0 + \epsilon_i \]

\[ = \beta_0 + \beta_1 x_i + \epsilon_i \]

**Women** (D=1):  
\[ y_i = \beta_0 + \beta_1 x_i + \beta_2 * 1 + \epsilon_i \]

\[ = \beta_0 + \beta_1 + \beta_2 + \epsilon_i \]

We see that the only difference in the equations for men and women is in the intercept – in which for men the intercept is \( \beta_0 \), for women it is \( \beta_0 + \beta_2 \). Thus when \( x=0 \), the expected salary for men is different from women. If the magnitude of \( \beta_2 \) is positive, then the expected salary boost for women is higher than men; if negative, then women receive on average a lower expected salary. Graphically, we expect to observe two parallel regression lines – lines that have the same slope of \( \beta_1 \) but intersect the y-axis at different points.

### 2.6 A checklist of things to do in Multiple Regression Analysis

You should consult the textbook for tips on writing up analyses of empirical projects. Here are a few others:

1. **Variable coding**: Check that all variables are coded correctly for missing observations and the intended response alternatives. Verify that direction of coding intuitively matches the variable (i.e. an ordinal measure of ‘education’ has higher scores for greater education).

2. **Theory in variables**: Capture your theory and plausible alternative explanations through variables; be prepared to include all in your regression.

3. **Causal ordering of variables and mediation**: Determine a good way to present the sensitivity of your results – think about the causal ordering of variables and how best to specify more parsimonious to more complete models.

4. **run the models** Run the Models. Estimate and re-estimate the regressions. Perform diagnostic tests; perform tests on blocks of coefficients to supplement sensitivity analysis.

5. **present results** Discuss results. Follow APSA style. And most importantly:

| If possible, avoid re-inventing the wheel! |
| Always find respected, quantitative studies on your topic to emulate. |
| Approach the regression analysis as they do, to the extent you are able! |