More on Regression

(8, 9, 10, and 11)

Tamer Çetin

Functional Form: *ln* Transformation

- Frequent nonlinear patterns better approximated with y or x transformed by taking **relative differences**:
- When transformed by taking the natural logarithm, differences in variable values *approximate relative differences*.
- Log differences work because differences in natural logs approximate percentage differences!

ln Transformation: Interpretation

- $\ln(x)$ = the natural logarithm of x
  - Sometimes we just say $\log x$ and mean $\ln(x)$.
- Log transformation allows for comparison in relative terms – percentages!
- Claim:

$$\ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$$

- The difference between the natural log of two numbers is approximately the relative difference between the two for small differences.

Interpreting parameters of regressions with log variables

- $\ln(y)^E = \alpha + \beta x_i$ - ‘log-level’ regression
- $y^E = \alpha + \beta \ln(x_i)$ - ‘level-log’ regression
- $\ln(y)^E = \alpha + \beta \ln(x_i)$ - ‘log-log’ regression
- Precise interpretation is key.
- The interpretation of the slope (and the intercept) coefficient(s) differs in each case!
- Often verbal comparison is made about a 10% difference in x if using level-log or log-log regression.

Comparing Different Models

Variables	**(1)** **price**	**(2)** **ln(price)**	**(3)** **price**	**(4)** **ln(price)**
Distance to city center, miles	-14.41	-0.13		
ln(distance to city center)			-24.77	-0.22
Constant	132.02	4.84	112.42	4.66
Observations	207	207	207	207
R-squared	0.157	0.205	0.280	0.334

Table: Hotel price and distance regressions

Multiple Linear Regression

- Multiple regression analysis uncovers average y as a function of more than one x variable: $y^E = f(x_1, x_2, \dots)$.
- It can lead to better predictions $\hat{y}$ by considering more explanatory variables.
- It may improve the interpretation of slope coefficients by comparing observations that are different in terms of one of the x_i variable but similar in terms of other x_{-i} variables ($-i$ means all other variable except i).
- Multiple linear regression specifies a linear function of the explanatory variables for the average y.

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \beta_k x_k$$

Multiple Linear Regression: Case of Two Regressors

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- β_1: the slope coefficient on x_1 shows difference in average y across observations with unit difference in x_1, *but the same value of* x_2.
- β_2 shows difference in average y across observations with with unit difference in x_2, *but the same value of* x_1.
- We can compare observations that are similar in one explanatory variable to see the differences related to the other explanatory variable.

Multiple Linear Regression: Case of Two Regressors

- If x_1 and x_2 are correlated, comparing observations with or without the same x_2 value makes a difference.
- If they are positively correlated, observations with higher x_2 tend to have higher x_1.
- In the simple regression we ignore differences in x_2 and compare observations with different values of x_1.
- But higher x_1 values mean higher x_2 values, too.
- Corresponding differences in y may be due to differences in x_1 but also differences in x_2.
- Neglecting x_2, when it is important leads to ‘omitted variable bias’.

Multiple Linear Regression: Case of Two Regressors

- Omitted variables are important, if you are interested in a coefficient value:
  - If you have a measure/variable on x_2 use it and you are done.
  - If you do not have a measure/variable on x_2:
    - similar to measurement errors: think and argue!
    - Is your 'true' parameter smaller or larger than what you estimated?
- Language: The slope on x_1 in the sample is confounded by omitting the x_2 variable, and thus x_2 is a **confounder.**
  - When you see/report coefficient values with adding more and more other variables to the model:
    - Want to show parameter stability - there is no other important confounder.
    - If your coefficient value changes by adding other variable(s), then you most likely have omitted variable bias problem.

Multiple Linear Regression: Some Language

- Multiple regression with two explanatory variables (x_1 and x_2),
- We condition on x_2, or control for x_2, when we include it in a multiple regression that focuses on average differences in y by x_1.
- We measure differences in expected y across observations that differ in x_1 but are similar in terms of x_2.
- Difference in y by x_1, *conditional on* x_2. OR, *controlling for* x_2.
- It is also called the controlled difference.

Collinearity

- **Perfectly collinearity** is when x_1 is a linear function of x_2.
- Consequence: cannot calculate coefficients (reason: linearly dependent matrix)
  - One will be dropped by software!
- Strong but imperfect correlation between explanatory variables is called *multicollinearity*.
  - Consequence: We can still get the slope coefficients and their standard errors, but:
    - Standard errors may be large.
    - Does not affect the value of β.

Using Qualitative Variables

- We can have qualitative variables as binary variables.
- Consider a qualitative variable like income categories or continents.
- How to add it to the regression model?
  - Create binary variables (dummy variables) for all options.
  - Add them - all but one. (Why? → linear dependence with the intercept!)
  - Left out one will be the base/reference!

Using Qualitative Variables

- x is a categorical variable with three values *low* , *medium* and *high.*
- Binary variable x_m denote if $x = medium$, x_h variable denote if $x = high$.
- $x = low$ is not included.
- It is called the *reference category* or left-out category.

$$y^E = \beta_0 + \beta_1 x_m + \beta_2 x_h$$

	x	x_medium	x_high
0	low	0	0
1	medium	1	0
2	high	0	1
3	low	0	0
4	high	0	1
5	medium	1	0

Using Qualitative Variables

$$y^E = \beta_0 + \beta_1 x_m + \beta_2 x_h$$

- β_0 shows average y in the reference category.
- Here, β_0 is average y when both $x_m = 0$ and $x_h = 0$:
  - this is the case of $x = low$.
- β_1 shows the difference of average y between observations with $x = medium$ and $x = low$.
- β_2 shows the difference of average y between observations with $x = high$ and $x = low$.

Interactions

- Interaction term refers to the product of two or more predictor variables.
- It is used to capture the combined effect or interaction between these variables on the response variable.
- To include an interaction term in a regression model:
  - Identify the predictor variables x_1 and x_2 that may have an interaction effect.
  - Create the interaction term by multiplying the two predictor variables together.
  - This can be done by adding a new column to the dataset representing the product of x_1 and x_2 as $x_1 * x_2$.
  - Fit the regression model by including the interaction term along with the main effects of the predictor variables.
  - The model equation is:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Interactions

- Where y^E is the response variable, $\beta_0, \beta_1, \beta_2,$ and β_3 are the regression coefficients, x_1 and x_2 are the predictor variables, $x_1 x_2$ is the interaction term, and ε is the error term.
- Estimate the coefficients in the regression model using a suitable regression technique such as ordinary least squares (OLS).
- The estimated coefficient β_3 represents the effect of the interaction term on the response variable, after controlling for the main effects x_1 and x_2.
- Adding interaction terms in a regression model, we analyze whether the relationship between y and x changes depending on another predictor.

Interactions

- Multiple regression offers the possibility to uncover such differences in patterns.
- For the simplest case, consider a regression with two explanatory variables: x_1 is quantitative; x_2 is binary.
- We wonder if the relationship between average y and x_1 is different for observations with $x_2 = 1$ than for $x_2 = 0$.
- Shall we uncover that difference?
- A multiple regression with x_1 and x_2 estimates two parallel lines for the $y - x_1$ pattern:
  - one for those with $x_2 = 0$ and one for those with $x_2 = 1$.

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Interactions

- If we want to allow for different slopes in the two x_2 groups, we have to do something different.
- That difference is including the interaction term.
- An **interaction term** is a new variable that is created from two other variables, by multiplying one by the other.
- In our case:

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- Not only are the intercepts different; the slopes are different, too:

$$y_0^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$
$$y_1^E = \beta_0 + \beta_2 + (\beta_1 + \beta_3) x_1$$

Interactions

- Interactions between right-hand-side variables in a linear regression allow for the slope coefficient of a variable to differ by values of another variable.
- Interactions between two right-hand-side variables are modeled in a linear regression as

$$y^E = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- β_1 shows average differences in y corresponding to a one-unit difference in x_1 when $x_2 = 0$.
- β_2 shows average differences in y corresponding to a one-unit difference in x_2 when $x_1 = 0$.
- β_3 is the coefficient on the interaction term.

Interactions

- It shows the additional average differences in y corresponding to a one-unit difference in x_1 when x_2 is one unit larger, too.
- When one of the two right-hand-side variables is binary, a simpler interpretation is also true.
- Say, $x_2 = 0$ *or* 1.
- Then,
    - β_1 shows the average difference in y corresponding to a one-unit difference in x_1 when $x_2 = 0$;
    - $\beta_1 + \beta_3$ shows the average difference in y corresponding to a one-unit difference in x_1 when $x_2 = 1$.

Modeling Probabilities: Linear Probability Model

- y represents an event:
  - whether something happens or not. $y = 0$ or 1.
- Conditional probabilities capture how the probability of such an event depends on the value of the conditioning variable(s).
- Thus, when we are interested in how the probability of the occurrence of the event depends on the values of x, we want to know $P[y = 1|x]$.
- Recall that the expected value of a 0–1 binary variable is also the probability that it is one.
- In other words, the average of a 0–1 variable is equal to the relative frequency of the value 1 among all observations.
- For example, an average of 0.5 corresponds to 50% of the observations being one; an average of 0.1 corresponds to 10% of ones.
- This is true whether the expectation, and thus the probability, is unconditional or conditional on some x variable(s).

Modeling Probabilities: Linear Probability Model

- In notation, when $y = 0$ or 1, then

$$E[y] = P[y = 1]$$

- $P[y = 1]$ represents the probability that y takes on the value 1. This probability can be interpreted as the proportion of cases where y equals 1 out of all possible cases.

$$E[y|x] = P[y = 1|x]$$

- This probability can be interpreted as the proportion of cases where y equals 1 out of all cases where x has the specific value.
- For this reason, the linear regressions with binary dependent variables show not only differences in expected y by x, but also differences in the probability of $y = 1$ by x.
- Because of this additional interpretation, a linear regression with a binary y is also called a **linear probability model** (LPM).
- A linear probability model with two or more explanatory variables is

$$y^P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

Modeling Probabilities: Linear Probability Model

- In other words, y^P denotes the probability that the dependent variable is one, conditional on the right-hand-side variables of the model.
- In effect, y^P is a shorthand replacing y^E.
- y^E would be fine here, too, but y^P conveys the message that y^E is not only a conditional expectation but also a conditional probability.
- In this model β_1 shows the difference in the probability that $y = 1$ for observations that are different in x_1 but are the same in terms of x_2 and all other right-hand-side variables.
- This interpretation is in addition to what's still true:
    - average difference in y corresponding to differences in x_1 with the other right hand-side variables being the same.

Modeling Probabilities: Linear Probability Model

- Just like any multiple linear regression, linear probability models allow for explanatory variables in logs, piecewise linear splines, polynomials, interactions, and so on.
- The interpretation of coefficients in such more complex models remains the same, with the added feature that differences in average y are also differences in the probability that $y = 1$.
- Similarly, all formulae and interpretations for standard errors, confidence intervals, hypotheses, and p-values of tests are the same.
- Linear probability models are not really different models:
  - they are the same linear regressions, only they allow for a specific interpretation beyond what's usual.
- However, because of the binary nature of y, the linear probability model has some issues.
- Less importantly, it is always heteroskedastic.
- Therefore, we should always estimate robust standard errors.
- The more important issue will concern the range of predicted values in an LPM, as we will discuss in the next section.

Predicted Probabilities in the Linear Probability Model

- Predicted values from linear probability models are **predicted probabilities**.
- Their interpretation is the estimated probability that $y = 1$ for observations with the particular value of the right-hand-side variables.
- Let's denote the predicted probability by $\hat{y}^P$.
- In the linear probability model with two or more explanatory variables, it is

$$\hat{y}^P = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots$$

- Let's pause a little bit to consider this terminology and notation.
- $\hat{y}^P$ is of course the usual predicted value from the linear regression.
- And, thus, it is the predicted mean of y, conditional on the values of the x variables.

Predicted Probabilities in the Linear Probability Model

- With quantitative y, this mean is what we would expect for an observation with those x values.
- However, y is binary here, so we can never expect its mean value for any observation.
- The mean value of y is a probability between zero and one; for any particular observation, the value of y is never between zero and one but is either zero or one.
- Hence the new term (predicted probability) and notation ($\hat{y}^P$).
- So predicted values from the linear probability model are predicted probabilities.
- And, therefore, they need to be between zero and one; they cannot be negative, and they cannot be greater than 1.

Predicted Probabilities in the Linear Probability Model

- Before illustrating predicting probabilities with our case study and moving on to other kinds of probability models, let's introduce the concept of **classification**.
- Classification is another kind of prediction with binary y variables.
- Instead of predicting the probability of $y = 1$ for target observations, the goal of classification is to put target observations into the $y = 1$ or $y = 0$ category.
- We can do that once a predicted probability is available:
  - we can put the observation in the $y = 1$ category if the predicted probability of $y = 1$ is high, and we can put it in the $y = 0$ category if the predicted probability is low.
  - What constitutes high and low is a decision to be made, with important consequences for the classification.
  - For this and other related reasons, classification has its own problems and solutions.

Non-linearity in Predicted Probabilities

- Polynomial Terms: One way to capture non-linear relationships is by including polynomial terms of the predictor variables in the regression model.
  - For example, you can add squared terms, cubic terms, or other higher-order terms to allow for curved or non-linear relationships between the predictors and the binary response variable.
- To circumvent the problem of predicting probabilities that are less than zero or greater than one, data analysts use two models as alternatives to the linear probability model:
  - these are called **logit** and **probit**.
- The linear probability model relates the probability of the $y = 1$ event to a linear combination of the explanatory variables.
- Logit and probit models relate the probability of the $y = 1$ event to a nonlinear function – called the **link function** – of the linear combination of the explanatory variables.

Non-linearity in Predicted Probabilities

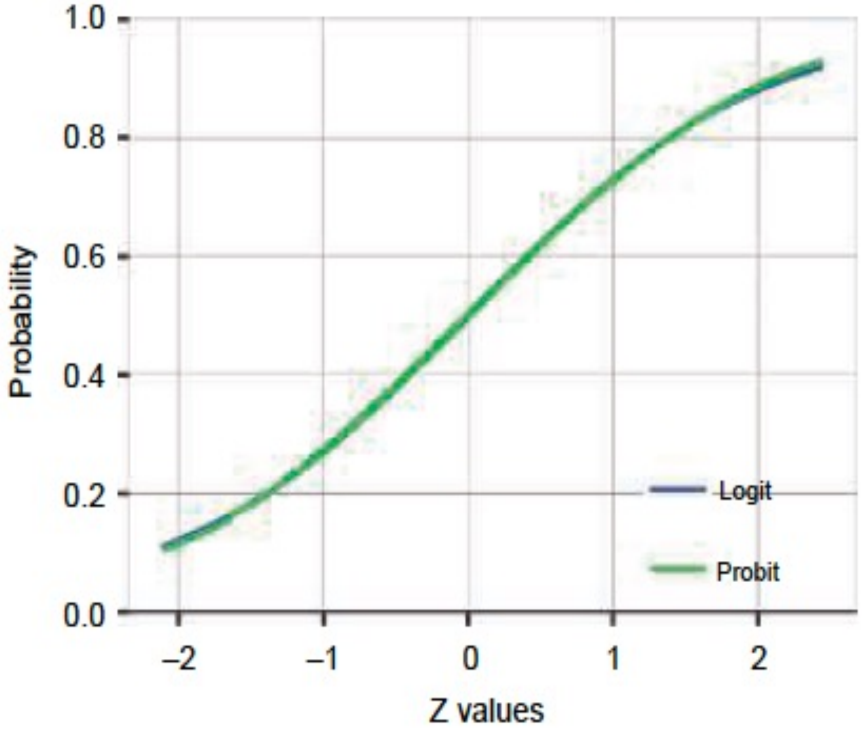
- The logit and the probit differ in the link function; however, both make sure that, in the end, the resulting probability is always strictly between zero and one.
- Both logit and probit models are parametric and assume specific functional forms.
- They do not explicitly model non-linear relationships between predictors and the response but rather focus on modeling the probability or odds of the binary outcome.

Non-linearity in Predicted Probabilities

- Both the probit and the logit transform the $\beta_0 + \beta_1 x_1 + \cdots$ linear combination using a link function that shows an S-shaped curve.
- Because of its shape, the slope of this link function is different at different places, depending on whatever is inside the link function.
- The slope is steepest when $y^P = 0.5$; it is flatter further away; and it becomes very flat if y^P is close to zero or one.
- All this implies that the difference in y^P that corresponds to a unit difference in any explanatory variable is not the same: it varies with the values of the x variables.
- This is in contrast with the linear probability model.

Non-linearity in Predicted Probabilities

- For the logit and probit, the same difference in an explanatory variable corresponds to a larger difference in y^P, the closer y^P is to 0.5.
- The same difference in an explanatory variable corresponds to a smaller difference in y^P, the closer y^P is to 0 or 1.
- All of this is a natural consequence of the S-shaped link that ensures that predicted probabilities are always between zero and one.



Non-linearity in Predicted Probabilities

- In a linear probability model, the coefficients have the usual interpretations.
- In logit and probit models the coefficients do not have the same interpretation.
- In fact, we do not interpret those coefficients.
- Instead, we transform them to arrive at an interpretation that is similar to what we have in linear probability models.
- To evaluate the magnitude of association between explanatory variables and y^P in logit and probit models, we compute and interpret **marginal differences**.
- The marginal difference corresponding to an explanatory variable is the average difference in y^P that corresponds to a one unit difference in that explanatory variable.
- Thus, marginal differences have the same interpretation as the coefficients of linear probability models.

Non-linearity in Predicted Probabilities

- Transformation of Variables: Non-linear relationships can also be captured by applying transformations to the predictor variables.
    - Common transformations include logarithmic, exponential, or power transformations.
    - These transformations can help capture non-linear patterns in the data and improve the fit of the model.
- EDA: Exploratory data analysis and model diagnostics can help in determining the most appropriate method to address non-linearity in your regression model with binary target variables.

Non-linearity in Predicted Probabilities

- Non-linear Machine Learning Algorithms: If the non-linear relationship between the predictors and the binary response variable is complex and cannot be adequately captured by the above methods, you can explore non-linear machine learning algorithms specifically designed for binary classification tasks.
    - These algorithms, such as decision trees, random forests, or neural networks, can capture complex interactions and non-linear relationships without explicitly specifying functional forms.

Interpretation of Statistics of Interest in Regression Analysis

- Probability models predict probabilities, which we denote by y^P.
- This is true for the linear probability model as well as for logit and probit models.
- Typically, these predicted probabilities take many values, somewhere between zero and one (or, for the linear probability model, sometimes beyond zero or one).
- In contrast, the dependent variable itself does not take on many values – just two: zero or one.
- As a result, probability models cannot fit zero–one dependent variables perfectly.
- Actually, they don't need to:
  - their task is to predict probabilities or uncover patterns of association between probabilities and the right-hand-side variables.
  - So, predicting values of a binary y would be a different task, called classification.

Interpretation of Statistics of Interest in Regression Analysis

- Goodness of fit of a regression tells us how good the estimated regression is in producing a prediction within the data.
- It is based on comparing actual y values with the predictions of an estimated model, and this comparison is done within the same dataset that we used for the estimation.
- When y is quantitative, we directly compare values of y to their predicted values from a regression.
- R-squared, the most widely used measure of fit for regressions, is the result of such a comparison, and so is the $\hat{y} - y$.

Interpretation of Statistics of Interest in Regression Analysis

- While the R-squared is a less natural measure of fit for probability models, we can calculate it just the same.
- Then we can use that R-squared to rank different models.

Interpretation of Statistics of Interest in Regression Analysis

- In regression models, standard errors, confidence intervals, hypotheses, and p-values provide valuable information about the statistical significance and precision of the estimated coefficients. Here are their interpretations:
- Standard Errors: Standard errors measure the variability or uncertainty associated with the estimated coefficients in the regression model.
  - A smaller standard error indicates a more precise estimate.
  - It represents the average amount by which the estimated coefficient would vary across different samples.
  - In general, smaller standard errors are desirable as they suggest more reliable estimates.
- Confidence Intervals: Confidence intervals provide a range of plausible values for the population parameter (e.g., regression coefficient) based on the sample data.
  - They are constructed around the estimated coefficient, typically at a chosen level of confidence (e.g., 95% confidence interval).
  - The interpretation is that if the same estimation were repeated multiple times and confidence intervals were constructed each time, approximately 95% of those intervals would contain the true population parameter.
  - A wider confidence interval indicates more uncertainty about the parameter estimate, while a narrower interval indicates more precision.

Interpretation of Statistics of Interest in Regression Analysis

- Hypotheses: In regression models, hypotheses are typically formulated to test whether a coefficient is statistically different from zero or to compare coefficients between different variables or groups.
  - For example, a hypothesis may state that the coefficient of a specific predictor variable is equal to zero (no effect).
  - Another hypothesis could test whether two coefficients are equal (e.g., comparing the effect of two different treatments).
  - Hypotheses are essential for testing specific research questions or theories within the regression framework.
- P-values: P-values quantify the strength of evidence against a null hypothesis.
  - They indicate the probability of observing the estimated coefficient (or a more extreme value) under the assumption that the null hypothesis is true.
  - A smaller p-value (typically below a pre-defined significance level, e.g., 0.05) suggests stronger evidence against the null hypothesis.
  - If the p-value is below the significance level, it is commonly interpreted as evidence to reject the null hypothesis in favor of an alternative hypothesis.

Interpretation of Statistics of Interest in Regression Analysis

- Overall, standard errors, confidence intervals, hypotheses, and p-values provide important information for assessing the statistical significance and reliability of the regression coefficients and for making inferences about the relationship between the predictors and the response variable in the population.