Multiple Regression Analysis

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$
- 4. Further Issues

Data Scaling and OLS Statistics

- We now return to the issue of changes in scale and origin we met before in Chapter 2 and examine the effects of rescaling the dependent or independent variables on *se*, *t* statistics, *F* statistics, and *CI*.
- As expected, when variables are rescaled, the coefficients, *se*, *CI*, *t* and *F* statistics change in ways that preserve all measured effects and testing outcomes.
- Hence, our conclusions are not affected by the units of measurement in the variables involved.

Data Scaling and OLS Statistics

- Consider the following estimated equation:

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

  and now consider what happens to our OLS statistics as we change the scale and origin of y and of x_1.

- We can work out these effects by simply manipulating the above equation.

Data Scaling and OLS Statistics

1. Changes in the scale of y: $c_1.y$

$$c_1.\hat{y} = (c_1.\hat{\beta}_0) + (c_1.\hat{\beta}_1)x_1 + (c_1.\hat{\beta}_2)x_2$$

- ✓ Coefficients are multiplied by c_1.
- ✓ Standard errors are multiplied by c_1.
- ✓ Statistical significance is not affected.
- ✓ *CI* change by the same factor, c_1.

Data Scaling and OLS Statistics

- Residuals are multiplied by c_1.
- SSR are multiplied by c_1^2.
- Standard Error of the Regression, $\mathrm{SER} = \hat{\sigma}$, is multiplied by c_1.
- R^2 is not affected, so the **overall significance of the regression** is not affected.

Data Scaling and OLS Statistics

2. Changes in the origin of y: $c_0 + y$

$$c_0 + \hat{y} = (c_0 + \hat{\beta}_0) + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- ✓ Only the intercept, β_0, is affected.
- ✓ The slope coefficients, measuring partial effects, are not affected.
- ✓ Residuals are not affected.
- ✓ R^2 is not affected.

Data Scaling and OLS Statistics

3. Changes in the scale of x_1: $d_1.x_1$

$$\hat{y} = \hat{\beta}_0 + (\hat{\beta}_1 / d_1)(d_1.x_1) + \hat{\beta}_2 x_2$$

- ✓ The coefficient associated to x_1, β_1, is divided by d_1.
- ✓ All other coefficients are not affected.
- ✓ The standard error of β_1 is divided by d_1.
- ✓ Statistical significance is not affected.
- ✓ The *CI* for β_1 change by the factor, $1/d_1$.

Data Scaling and OLS Statistics

- Residuals are not affected.
- Hence, neither SSR nor the SER are affected.
- R^2 is not affected, so the **overall significance of the regression** is not affected.

Data Scaling and OLS Statistics

4. Changes in the origin of x_1: $d_0 + x_1$

$$\hat{y} = (\hat{\beta}_0 - \hat{\beta}_1 d_0) + \hat{\beta}_1 (x_1 + d_0) + \hat{\beta}_2 x_2$$

- Only the intercept, β_0, is affected.
- The slope coefficients, measuring partial effects, are not affected.
- Residuals are not affected.
- R^2 is not affected.

Data Scaling and OLS Statistics

- **Conclusion**: Changes in scale and/or origin does not affect to any substantial part of the regression.
- In particular, statistical significance and interpretation of coefficients is not affected by data scaling.
- Note that to make our equation invariant to the origin of the variables we need an intercept in our equation.

Data Scaling and OLS Statistics

- This analysis shows clearly that if variables appear in logarithmic form, changing the units of measurement does not affect the slope coefficients.
- This follows from the fact that

$$\log(c_1 . y) = \log(c_1) + \log(y) \qquad c_1 > 0$$

$$\log(d_1 . x_j) = \log(d_1) + \log(x_j) \qquad d_1 > 0$$

so only the intercept is affected in these cases.

Beta Coefficients

- Sometimes in econometric applications, a key variable is measured on a scale that is difficult to interpret, for example, test scores, synthetic indexes,…
- In such cases, we can be interested in see what happens to y when the corresponding independent variable varies by one standard deviation.

Beta Coefficients

- Sometimes, it is useful to obtain regression results when *all* variables involved, *y* as well as the *x*´s, have been *standardized*.
- To standardize a variable subtracts its mean and divide by its standard deviation.
- Why is standardization useful?

  Lets see what this transformation implies for the coefficient estimates.

Beta Coefficients

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} + \hat{u}_i$$

Averaging this equation and subtracting

$$y_i - \bar{y} = \hat{\beta}_1 (x_{i1} - \bar{x}_1) + \hat{\beta}_2 (x_{i2} - \bar{x}_2) + \ldots + \hat{\beta}_k (x_{ik} - \bar{x}_k) + \hat{u}_i$$

Simple algebra gives us the estimated equation in standardized form

$$\frac{y_i - \bar{y}}{\hat{\sigma}_y} = \left(\hat{\beta}_1 \frac{\hat{\sigma}_1}{\hat{\sigma}_y}\right)\left[\frac{(x_{i1} - \bar{x}_1)}{\hat{\sigma}_1}\right] + \left(\hat{\beta}_2 \frac{\hat{\sigma}_2}{\hat{\sigma}_y}\right)\left[\frac{(x_{i2} - \bar{x}_2)}{\hat{\sigma}_2}\right] + \ldots + \left(\hat{\beta}_k \frac{\hat{\sigma}_k}{\hat{\sigma}_y}\right)\left[\frac{(x_{ik} - \bar{x}_k)}{\hat{\sigma}_k}\right] + \frac{\hat{u}_i}{\hat{\sigma}_y}$$

Beta Coefficients

Which we can rewrite as

$$z_y = \hat{b}_1 z_1 + \hat{b}_2 z_2 + \ldots + \hat{b}_k z_k + \hat{e}$$

where z denotes an standardized variable, the z-score, $\hat{e}$ denotes the error and the new coefficients are

$$\hat{b}_j = \hat{\beta}_j \frac{\hat{\sigma}_j}{\hat{\sigma}_y} \qquad \text{for } j = 1, 2, \ldots, k$$

These $\hat{b}_j$ are traditionally called **standardized coefficients** or **beta coefficients**.

Beta Coefficients

- The meaning of these coefficients is as follows: If x_j increases by one standard deviation, then $\hat{y}$ changes by $\hat{b}_j$ standard deviations, holding all other variables constant.
- Thus, we are measuring effects not in terms of the original units of y and x_j, but in standard deviation units.
- Because the equation in terms of the z-score makes the scale of the regressors irrelevant, this equation puts the explanatory variables on equal footing.

Beta Coefficients

- In a standard OLS equation, it is not possible to simply look at the size of different coefficients and conclude that the explanatory variable with the largest coefficient is "the most important".
- We just have seen that the magnitudes of coefficients can be changed at will by changing the scale of x_j.
- But, when each x_j has been standardized, comparing magnitudes of the resulting beta coefficients is more compelling.

Functional Form

- OLS can be used for modeling relationships that are not strictly linear in x and y by using nonlinear functions of x and y, if the model is still linear in the parameters.
- We consider some possibilities that often appear in applied work:

1. log´s of x and y.
2. quadratic forms of x.
3. Interactions of x variables.

Proportions and Percentages

- Remember that:

1. Proportional change: $$\frac{x_1 - x_0}{x_0} = \frac{\Delta x}{x_0}$$

2. Percentage change: $$100.\frac{\Delta x}{x_0} = \%\Delta x$$

3. Elasticity: $$\frac{\Delta y}{\Delta x}.\frac{x_0}{y_0} = \frac{\%\Delta y}{\%\Delta x}$$

Proportions and Percentages

4. Changes in logarithms:

$$\Delta \log(x) = \log(x_1) - \log(x_0) \approx \frac{x_1 - x_0}{x_0} = \frac{\Delta x}{x_0}$$

Hence,

$$100.\Delta \log(x) \approx \%.\Delta x$$

A Linear Model for log(y)

- Consider the model

$$\log(y) = \beta_0 + \beta_1 x + u$$

- What is the meaning of β_1 in this model?
- If $\Delta u = 0$, then x has a linear effect on log(y):

$$\Delta \log(y) = \beta_1 \Delta x$$

  or,

$$\%\Delta y = (100.\beta_1).\Delta x$$

  i.e. $100.\beta_1$ is the percentage change in y by unit of x.

A Constant Elasticity Model

- Consider the model

$$\log(y) = \beta_0 + \beta_1 \log(x) + u$$

- What is the meaning of β_1 in this model?
- If $\Delta u = 0$, then $\log(x)$ has a linear effect on $\log(y)$:

$$\Delta\log(y) = \beta_1 \Delta\log(x) \quad \Leftrightarrow \quad \%\Delta y = \beta_1 . \%\Delta x$$

i.e. β_1 is the elasticity of y with respect to x.

Functional Forms Involving logs

Model	Dependent Variable	Independent Variable	Interpretation of β_1
level-level	y	x	$\Delta y = \beta_1 . \Delta x$
level-log	y	$\log(x)$	$\Delta y = (\beta_1/100) . \% \Delta x$
log-level	$\log(y)$	x	$\% \Delta y = (100 . \beta_1) . \Delta x$
log-log	$\log(y)$	$\log(x)$	$\% \Delta y = \beta_1 . \% \Delta x$

Functional Form

- **Important**: While the mechanics of the linear regression does not depend on how *y* and the *x*´s are defined, the interpretation of the coefficients does depend on their definitions.

Why use log models?

- Using log´s leads to coefficients with appealing interpretations, i.e. elasticity or semi-elasticity.
- Models with log´s are invariant to the scale of the variables, since they measure proportional changes.
- For models with $y > 0$, using $\log(y)$ as the dependent variable often satisfy the CLM assumptions more closely than models using the level of y.
- For models with $y > 0$, the conditional distribution is often heteroskedastic or skewed, while $\log(y)$ is much less so.

Why use log models?

- Taking log´s usually narrows the range of the variable. This makes estimates less sensitive to outlying (or extreme) observations on the dependent or independent variables.
- One limitation of the log is that it can not be used if a variables can take zero or negative values.
- One drawback to using a dependent variable in log form is that it is more difficult to predict the original variable. The original model allows us to predict log(*y*), not *y*.

Why use log models?

- Also it is *not* legitimate to compare R^2 from models where y is the dependent variable in one case and log(y) is the dependent variable in the other. These measures explained variations in different variables.

- **Important**: This is a general rule, the R^2 cannot be used to compare models with different dependent variable.

Some Rules of Thumb

- What types of variables are often used in log form?
  - ✓ Variables in money terms that must be positive.
  - ✓ Very large variables, such as population.

- What types of variables are often used in level form?
  - ✓ Variables measured in years.
  - ✓ Variables that are a proportion or percent, i.e. inflation, interest rates.

Quadratic Models

- A quadratic model is of the form

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

- **Quadratic functions** are also used quite often in applied economics to capture decreasing or increasing marginal effects.
- **Important**: β_1 does not measure the change in y with respect to x; it makes no sense to hold x^2 fixed while changing x.

Quadratic Models

- If $\Delta u = 0$ then,

$$\Delta y \approx (\beta_1 + 2\beta_2 x).\Delta x \quad \Rightarrow \quad \frac{\Delta y}{\Delta x} \approx \beta_1 + 2\beta_2 x$$

the marginal effect of x on y depends linearly on the value of x.

The estimated slope is $\beta_1 + 2\beta_2 x$.

- In a particular application this marginal effect should be evaluated at interesting values of x.

More on Quadratic Models

- Suppose that $\beta_1 > 0$ and $\beta_2 < 0$.
- Then *y* is increasing in *x* at first, but will eventually turn around and be decreasing in *x*.
- The turning point will be at

$$x^* = \left| \frac{\beta_1}{2\beta_2} \right|$$

More on Quadratic Models

- Suppose that $\beta_1 < 0$ and $\beta_2 > 0$.
- Then y is decreasing in x at first, but will eventually turn around and be increasing in x.
- The turning point will be at

$$x^* = \left| \frac{\beta_1}{2\beta_2} \right|$$

which is the same as before.

Interaction Terms

- Sometimes, it is natural for the partial effect, elasticity or semi-elasticity of the dependent variable with respect to an explanatory variable to depend on the magnitude of yet another explanatory variable.
- These effects can be modeled through **interaction terms**, $x_i x_j$.

Interaction Terms

- Consider the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

- In this case β_1 is not the partial effect of x_1 on y, because there is an **interaction term**, $x_1 x_2$.
- If $\Delta u = 0$ then,

$$\Delta y = (\beta_1 + \beta_3 x_2).\Delta x_1 \quad \Rightarrow \quad \frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2$$

Interaction Terms

- The partial effect of x_1 on y depends linearly on x_2.
- In summarizing the effect of x_1 on y, we must evaluate the above expression at interesting and representative values of x_2, for examples the sample mean of x_2.

Functional Form

- This shows clearly that the partial effects of x_j on y are constant only if the model is linear in variables. In all other cases the interpretation of the coefficients does depend on the definitions of the variables.

R-Squared

- We found before the R^2 as a **goodness of fit** measure.
- R^2 is simply an estimate of how much variation in y is explained by the $\boldsymbol{x}$´s, and even it is intuitively obvious that a higher R^2 is preferable to a lower one, nothing about the classical model assumptions requires that R^2 be above any particular value.
- A small R^2 does imply that the error variance is large relative to the variance of y, which means that the β_j are not precisely estimated.

R-Squared

- But remember, that a large error variance can be offset by a large sample size, so if n is large enough, we may be able to precisely estimate the partial effects even though we have not controlled for many unobserved factors.
- Also that the relative *change* in the R^2, when variables are added to an equation, is very useful: the F statistic for testing the joint significance of the added variables crucially depends on the difference in the R^2 between the unrestricted and the restricted models.

Adjusted *R*-Squared

- Recall that the R^2 will always increase as more variables are added to a given model.
- This can lead to the false impression that models with more explanatory variables are always preferred, but this is completely false. If we add variables to a given model, R^2 will never decrease, even if these variables are not significant.
- To avoid this algebraic fact we can "adjust" the R^2 in a way that takes into account the number of variables included in a given the model.

Adjusted *R*-Squared

- To see how the usual R^2 might be adjusted, it is usefully written as

$$R^2 = 1 - \frac{\mathrm{SSR}/n}{\mathrm{SST}/n}$$

- This expression reveals what R^2 is actually estimating.
- The **population** R^2 is defined as $1 - \dfrac{\sigma_u^2}{\sigma_y^2}$

Adjusted R-Squared

- This is what R^2 is supposed to be estimating.
- However, we have better estimates for these variances that the ones used in the R^2. So lets use unbiased estimates for these variances

$$\bar{R}^2 = 1 - \frac{\mathrm{SSR}/(n-k-1)}{\mathrm{SST}/(n-1)} = 1 - \left(1 - R^2\right)\frac{n-1}{n-k-1}$$

- This is the *adjusted* R^2.

Adjusted *R*-Squared

- The primary attractiveness of $\bar{R}^2$ is that it imposes a penalty for adding additional independent variables to a model.
- If an independent variable is added to a model then SSR falls, but so does the *df* in the regression, $n - k - 1$. So $\bar{R}^2$ can go up or down when a new independent variable is added to a regression.

Adjusted *R*-Squared

- An interesting algebraic fact is that if we add a new independent variable to a regression equation, $\bar{R}^2$ increases if, an only if, the t statistic on the new variable is greater than one in absolute value.
- Thus we see immediately that using $\bar{R}^2$ to decide whether a certain independent variable belongs in a model gives us a different answer than standard t testing.

Goodness of Fit

- It is important not to focus too much on R^2 or $\bar{R}^2$, and lose insights from economic theory and common sense.
- Goodness of fit by itself is not an objective.
- If economic theory clearly predicts a variable belongs to a model, generally leave it in.
- Don't try to include a variable that prohibits a sensible interpretation of the variables of interest. Remember the *ceteris paribus* interpretation of multiple regression.

Goodness of Fit

- Provided the above conditions are fulfilled, you can use the R^2 to measure the goodness of fit of models with the same number of independent variables and the same y:

(1) $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$

(2) $y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + u$

- These are **nonnested models**, because neither equation is a special case of the other.

Goodness of Fit

- You can use the $\bar{R}^2$ to measure the goodness of fit of models with different number of independent variables **and** the same *y*:

$$(1) \qquad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

$$(2) \qquad y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \beta_4 \log(x_4) + u$$

- Explanatory variables can appear with different functional form, but not *y*.

Goodness of Fit

- You **cannot** use neither the R^2 nor $\bar{R}^2$ to measure the goodness of fit of models with different functional forms for the dependent variable, y:

(1) $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

(2) $$\log(y) = \beta_0 + \beta_1 x_1 + \beta_4 \log(x_4) + u$$

- The reason is simple: the variation to be explained, SST, is different for both models.

Prediction

- Suppose we have estimated the equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

- When we plug in particular values of the ***x*´s**, we obtain a prediction for *y*, which is an estimate of the expected value of *y* given the particular values for the ***x*´s**.
- Let $c_1, c_2, \ldots, c_k$ denote the particular values for each of the *k* independent variables; these may or may no correspond to an actual data point in our sample.

Prediction

- The parameter we would like to estimate is

$$\theta_0 = E(y \mid x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k)$$
$$= \beta_0 + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_k c_k$$

- The natural estimator of θ_0 is

$$\hat{\theta}_0 = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \hat{\beta}_2 c_2 + \ldots + \hat{\beta}_k c_k$$

- This is easy to compute once the model has been estimated.
- **Predictions** are certainly useful, but they are subject to sampling variation, so what about its uncertainty?

Prediction

- It is natural to construct a confidence interval for θ_0 which is centered at $\hat{\theta}_0$.
- To obtain a *CI* for θ_0, we need a standard error for $\hat{\theta}_0$
- Then, under MLR6 we can construct a 95% *CI* as $\hat{\theta}_0 \pm t_{.025}.se(\hat{\theta}_0)$, where $t_{.025}$ is the 97.5th percentile in the t_{n-k-1} distribution.
- Otherwise, with a large *df*, we can construct a 95% *CI* using the *rule of thumb* $\hat{\theta}_0 \pm 2.se(\hat{\theta}_0)$, since for large $n-k-1$ then $t_{.025} \approx 1.96$

Prediction

- How do we obtain the *se* of $\hat{\theta}_0$?
- If the computer software does not do the job for you, note that all you need is a *se* of a linear combination of the OLS estimators, just as in hypothesis testing, so the same trick we used there works here.
- Write $\beta_0 = \theta_0 - \beta_1 c_1 - \beta_2 c_2 - \ldots - \beta_k c_k$, and plug this into the equation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

to obtain

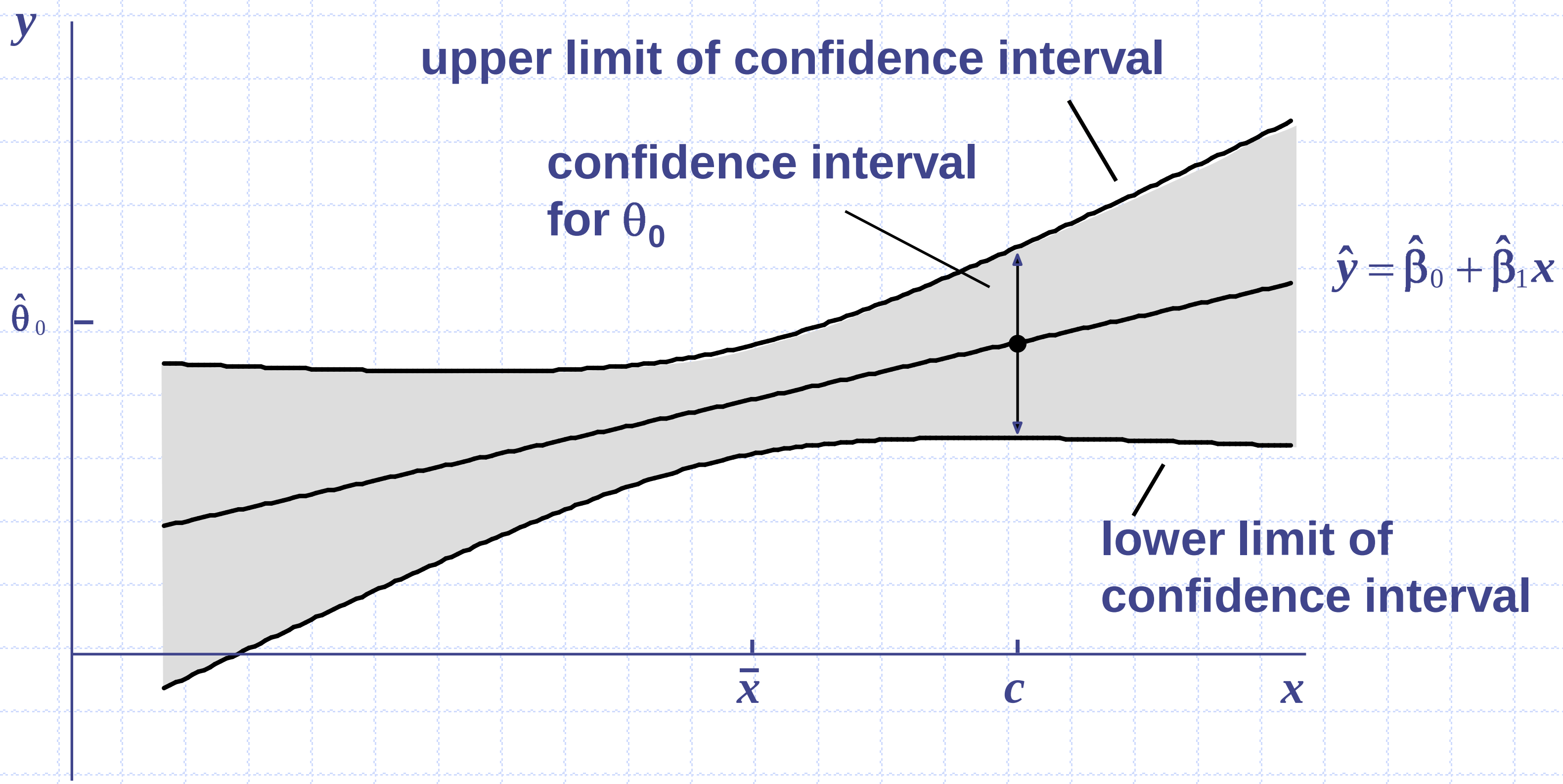
$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + \ldots + \beta_k(x_k - c_k) + u$$

Prediction

- In other words, we subtract the value c_j from each observation on x_j, and then we run the regression of

  y_i on $(x_{i1} - c_1), (x_{i2} - c_2), \ldots , (x_{ik} - c_k), \ \ i = 1,\ldots,n$
- The predicted value, and more importantly, its *se*, are obtained from the *intercept*, or constant, in this regression.
- Note that the *se* will be smallest when the *c*´s are equal to the mean of the *x*´s.
- This result is not surprising, since intuitively we have less uncertainty near the middle of our data.

Prediction: *CI*



This illustrates graphically the confidence interval for predictions in the SLR case.

Prediction

- The previous method allows us to put a *CI* around the OLS estimate of $E(y|x_1,x_2,\ldots,x_3)$, for any values of the ***x***´s.
- In other words, we obtain a *CI* for the average value of *y* for the subpopulation with a given set of covariates.
- But **a *CI* for the average unit** in the subpopulation **is not exactly the same as a *CI* for a particular unit** in the subpopulation.
- In forming a *CI* for an unknown outcome on *y*, we must account for another very important source of variation: the variance in the unobserved error, which measures our ignorance on the unobserved factors that affect *y*.

Prediction Interval

- Let y^0 denote the value for which we would like to construct a *CI*, usually called **prediction interval**. Let $x_1^0, x_2^0, \ldots, x_k^0$ be the new values of the ***x***´s, which we observe, and let u^0 be the unobserved error. Therefore, we have

$$y^0 = \beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \ldots + \beta_k x_k^0 + u^0$$

- As before, our best point prediction of y^0 is the expected value of y^0 given the explanatory variables, which we estimate from the OLS regression line

$$\hat{y}^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \hat{\beta}_2 x_2^0 + \ldots + \hat{\beta}_k x_k^0$$

Prediction Interval

- The **prediction error** in using $\hat{y}^0$ to predict y^0 is

$$\hat{e}^0 = y^0 - \hat{y}^0 = (\beta_0 + \beta_1 x_1^0 + \beta_2 x_2^0 + \ldots + \beta_k x_k^0) + u^0 - \hat{y}^0$$

  Because OLS estimators are unbiased and $E(u^0) = 0$, then $E(\hat{e}^0) = 0$. So the expected prediction error is zero.
- In finding the variance of $\hat{e}^0$, note that u^0 is uncorrelated with $\hat{y}^0$ (why?).
- Therefore, the **variance of the prediction error** (conditional on the ***x***´s) is the sum of the variances

$$Var(\hat{e}^0) = Var(\hat{y}^0) + Var(u^0) = Var(\hat{y}^0) + \sigma^2$$

Prediction Interval

- There are two sources of variation in $\hat{e}^0$.

1. The sampling error in $\hat{y}^0$, which arises because we have estimated the β_j.
2. The ignorance of the unobserved factors that affect y, which is reflected in σ^2.

- Under the CLM assumptions $\hat{e}^0$ is also normally distributed (conditional on the $\mathbf{x}$´s). And using unbiased estimators of $Var(\hat{y}^0)$ and σ^2, we can define the *se* of $\hat{e}^0$ as

$$se(\hat{e}^0) = \left\{ \left[se(\hat{y}^0) \right]^2 + \hat{\sigma}^2 \right\}^{\frac{1}{2}}$$

Prediction Interval

- Using the same reasoning for the t statistic of the $\hat{\beta}_j$, $\dfrac{\hat{e}^0}{se(\hat{e}^0)}$ has a t distribution with n–k–1 df. Therefore,

$$\Pr\left[-t_{.025} \leq \frac{\hat{e}^0}{se(\hat{e}^0)} \leq t_{.025}\right] = .95$$

  where $t_{.025}$ is the 97.5th percentile in the t_{n-k-1} distribution.
- Plugging in $\hat{e}^0 = y^0 - \hat{y}^0$ and rearranging gives a 95% **prediction interval** for y^0: $\hat{y}^0 \pm t_{.025} . se(\hat{e}^0)$.

Prediction Interval

- Usually the estimate of σ^2 is much larger than the variance of the prediction.
- Thus, this prediction interval will be much wider than the simple *CI* for the prediction.

- As before with a large *df*, we can construct a 95% prediction interval using the *rule of thumb* $\hat{y}^0 \pm 2.se(\hat{e}^0)$, since for large $n\text{–}k\text{–}1$ then $t_{.025} \approx 1.96$.

Residual Analysis

- Sometimes, it is useful to examine the residuals for the individual observations. This process is known as **residual analysis**.
- Big residuals, either positive or negative, can be informative about special events or characteristics of individual observations.
- Extreme residuals, greater in absolute value than 3 standard error of the regression, are called **outliers**.
- Outliers merit some consideration since they can influence estimation results.

Predicting *y* in a log(*y*) model

- Define *logy* = log(*y*), and consider the problem of predicting *y* when the estimated model is

$$logy = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- Given OLS estimators we predict *logy* as

$$\hat{logy} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

- Simple exponentiation, $\hat{y} = \exp(\hat{logy})$, will systematically *underestimate* the expected value of *y*.
- Instead, we need to scale this up by an estimate of the expected value of exp(*u*).

Predicting *y* in a log(*y*) model

- Note that if $u \sim N(0,\sigma^2)$, then $E(\exp(u)) = \exp\left(\sigma^2/2\right)$
- Under the CLM assumptions MLR.1 through MLR.6, then

$$E(y|\mathbf{x}) = \exp(\sigma^2/2).\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

- This equation shows that, under normality, the simple adjustment needed to predict *y* is

$$\hat{y} = \exp(\hat{\sigma}^2/2).\exp(\widehat{logy})$$

where $\hat{\sigma}^2$ is the unbiased estimator of σ^2.

- Because $\hat{\sigma}^2 > 0 \quad \Rightarrow \quad \exp(\hat{\sigma}^2/2) > 1$

Predicting *y* in a log(*y*) model

- The above prediction is not unbiased, but it is consistent. And in many cases works pretty well.
- However, it does rely on the normality of *u*.
- It is useful to have a prediction that does no rely on normality. If we just assume that *u* is independent of the $\boldsymbol{x}$´s, then we have

$$E(y\,|\,\boldsymbol{x}) = \alpha_0.\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

where α_0 is the expected value of exp(*u*), which must be greater than unity.

Predicting *y* in a log(*y*) model

- Given an estimate $\hat{\alpha}_0$, we can predict *y* as

$$\hat{y} = \hat{\alpha}_0.\exp(l\hat{o}gy)$$

- It turns out that a consistent estimator of $\hat{\alpha}_0$ is easily obtained:

1. Obtain the fitted values $l\hat{o}gy_i$
2. Create $\hat{m}_i = \exp(l\hat{o}gy_i)$
3. Regress *y* on $\hat{m}$, *without* an intercept. The coefficient on $\hat{m}$, the only coefficient there is, is the estimate of α_0, i.e. $E(\exp(u))$.
4. Once $\hat{\alpha}_0$ is obtained, predict *y* as $\hat{y} = \hat{\alpha}_0.\exp(l\hat{o}gy)$.

Comparing log(y) and y models

- As mentioned before, R^2 cannot be used to compare models with different dependent variables. In particular, it cannot be used to compare models with y and log(y) as dependent variables.
- If the goal is to find a goodness-of-fit measure in the log(y) model that can be compared with the R^2 from a model where y is the dependent variable we can use the previous results.
- After running the regression of y on $\hat{m}$ through the origin, we obtain the fitted values for this regression, $\hat{y}_i = \hat{\alpha}_0 . \hat{m}_i$

Comparing log(y) and y models

- Then, we find the sample correlation between $\hat{y}_i$ and the actual y_i in the sample.
- The *square* of this *can* be compared with the R^2 we get by using y as the dependent variable in a linear regression model.

- Remember that the R^2 in the fitted equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

is just the squared correlation between y_i and $\hat{y}_i$.