ECON5529
Bayesian Theory

Lecture 6

Ellis Scharfenaker

Regression models

- Most economic studies concern the relationship between two or more variables. In general, interest centers on the conditional distribution of y given x, parameterized by θ as $p[y \mid x, \theta]$, under a model in which the n observations $\{x, y\}_i$ are *exchangeable*.
- Regression models are from the class of models in which the data, (*response* or *outcome variable*) $y = \{y_1, \ldots, y_n\}$, is to be explained by some *auxiliary variable* or *covariates*.
- They consist of a *deterministic model* of the data $y[x]$ conditional on explanatory variables (*treatment* or *control variables*) $x = \{x_1, \ldots, x_k\}$, together with an *error model* assigning a likelihood to any deterministic model that depends on the size of the residual errors generated from that model applied to the actual data, $e_j = y_j - y[x_j]$.
- An intuitive way to understand the regression idea is to view a particular model as a deterministic *prediction* of y_i given x_i. Economists usually express this as a functional relationship determined by some parameters θ, which constitutes the hypothesis:

  $y_i = f[x_i; \theta]$
- Taylor's Rule from calculus tells us that any sufficiently differentiable function can be approximated by a polynomial near a particular value of its argument:

```
Series[y[x], {x, 0, 2}]
```

  $y[0] + y'[0]\, x + \frac{1}{2} y''[0]\, x^2 + O[x]^3$
- Factoring the second order Taylor expansion:

  $y[x] = y[0] + (y'[0] + y''[0]\, x)\, x$
- Let $\beta_0 = y[0]$ and $\beta_1 = (y'[0] + y''[0]\, x)$

  $y[x] = \beta_0 + \beta_1\, x$
- We can see the response of the function $y[x]$ to a change in x depends on x itself, since it is equal to $y'[0] + y''[0]\, x$.
- A ***linear deterministic model*** assumes that $y'' = 0$, so the response of y to x is the same for all relevant values of x (the *domain* of the function).
- This implies that the linear regression model essentially averages all of the correlations of y and x over the whole domain in trying to fit a linear deterministic model of the form:

  $y_i = \beta_0 + \beta_1\, x_i$

- Of course, real data almost never lies exactly on the straight line that the deterministic model predicts. Therefore, if we take a regression model literally, the likelihood of the data is going to be zero.
- In order to make regression models consistent with real data, we introduce the possibility of errors, and some assumption about the prior probability of errors, so that the likelihood function will assign a non-zero probability to the data conditional on the hypothesized model.

  $y_i = \beta_0 + \beta_1 x_i + e_i$, or

  $e_i = y_i - (\beta_0 + \beta_1 x_i)$

Basic Linear regression

- Suppose the model class we are interested in is the set of multiple linear relations $\boldsymbol{y}[\boldsymbol{X}; \boldsymbol{\beta}] = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{X}$ is an $n \times k$ *design matrix* of explanatory variables with a leading vector of ones for the constant, $\boldsymbol{\beta}$ is a $k \times 1$ vector of coefficients to be estimated, $\boldsymbol{y}$ is an $n \times 1$ vector of outcome variable values distributed $\mathcal{N}[\boldsymbol{X\beta}, \sigma^2 I]$, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of errors.

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}, \ \boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & .. & .. & x_{1k} \\ x_{21} & x_{22} & .. & .. & .. \\ .. & .. & .. & .. & .. \\ .. & .. & .. & .. & .. \\ x_{n1} & x_{n2} & .. & .. & x_{nk} \end{pmatrix}, \ \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{pmatrix}, \ \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{pmatrix}$$

- The simplest version of the linear regression model is to model the elements of $\boldsymbol{\epsilon}$ as exchangeable and distributed $\mathcal{N}[0, \sigma^2 I]$ for a constant σ^2 where $\boldsymbol{\beta}$ is assumed to be *independent* of σ^2.
- Thus the likelihood function for a sample size of n is:

  $$p\left[\boldsymbol{\epsilon} \,\middle|\, \sigma^2\right] = \left(2\,\pi\,\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}\left(\boldsymbol{\epsilon}^T \boldsymbol{\epsilon}\right)}$$

- Where by use of matrix notation $\boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = \Sigma_{i=1}^{n} \epsilon_i^2$
- Using the simple linear model this implies for the transformed error term:

  $$p\left[\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2\right] = \left(2\,\pi\,\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}\left((\boldsymbol{y}-\boldsymbol{X\beta})^T (\boldsymbol{y}-\boldsymbol{X\beta})\right)}$$

- We know from any basic econometrics course the values for which $p\left[\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2\right]$ is maximized:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}' \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \ \text{ and } \ \hat{\sigma}^2 = \left(\frac{\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^T \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)}{n-k} \right)$$

- So far we have the standard frequentist approach to linear modeling and once the data $(\boldsymbol{X}, \boldsymbol{y})$ are observed, the likelihood function is maximized relative to the unknown parameter vector $\boldsymbol{\beta}$ and the unknown scalar σ.
- However, we are really interested in the joint posterior distribution:

  $$p\left[\boldsymbol{\beta}, \sigma^2 \,\middle|\, \boldsymbol{X}, \boldsymbol{y}\right] \propto p\left[\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2\right] p[\boldsymbol{\beta}]\, p\left[\sigma^2\right]$$

- From which we can calculate the marginal distributions for $\boldsymbol{\beta}$ and σ^2.

Uninformative Priors for the Linear Model

- Using the standard ignorance priors for the normal distribution $p[\boldsymbol{\beta}] \propto d\boldsymbol{\beta} = c$ over the support $(-\infty,\infty)$ and $p[\sigma^2] = \frac{1}{\sigma}$ over $(0, \infty)$
- Assuming independence of $p[\boldsymbol{\beta}]$ and $p[\sigma^2]$ such that $p[\boldsymbol{\beta}, \sigma^2] = p[\boldsymbol{\beta}]\, p[\sigma^2]$ we can derive the joint posterior:

$$p[\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{X}, \boldsymbol{y}] \propto p[\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2]\, p[\boldsymbol{\beta}]\, p[\sigma^2]$$

$$= \left(2\pi\sigma^2\right)^{-n/2} e^{-\frac{1}{2\sigma^2}((\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}))} \frac{1}{\sigma}$$

$$\propto \sigma^{-n-1} e^{-\frac{1}{2\sigma^2}\left(\hat{\sigma}^2(n-k)+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T \boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)}$$

- Proof

$$p[\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2] = \left(2\pi\sigma^2\right)^{-n/2} \text{Exp}\left[-\frac{1}{2\sigma^2}((\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}))\right]\frac{1}{\sigma}$$

$$\propto \sigma^{-n-1} \text{Exp}\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta})\right]$$

$$= \sigma^{-n-1} \text{Exp}\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})'\boldsymbol{X}'\boldsymbol{y} + 2((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})'\boldsymbol{X}'\boldsymbol{X}((\boldsymbol{X}'\boldsymbol{X}))^{-1}\boldsymbol{X}'\boldsymbol{y})\right]$$

$$= \sigma^{-n-1} \text{Exp}\left[-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}\right) + \hat{\boldsymbol{\beta}}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}\right]$$

$$= \sigma^{-n-1} e^{-\frac{1}{2\sigma^2}\left(\hat{\sigma}^2(n-k)+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)'\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)}$$

Marginal posterior distribution of β

- To derive the marginal posterior distribution $p[\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}]$ make the transformation $\tau = \sigma^{-2}$ and integrate with respect to τ

$$p[\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}] = \int_0^\infty p[\boldsymbol{X}, \boldsymbol{y} \mid \beta, \sigma^2]\, p[\beta]\, p[\sigma^2]\, d\tau$$

$$= \int_0^\infty \tau^{\frac{n}{2}-1} e^{-\frac{1}{2}\tau\left(\hat{\sigma}^2(n-k)+\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T \boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right)\right)} d\tau$$

- Recognizing the integral as the kernel of a Gamma distribution:

$$\text{Gamma}[\alpha, \lambda] = \frac{\lambda^{\alpha+1}}{\Gamma[\alpha+1]} \tau^\alpha e^{-\lambda\tau} \propto s^\alpha e^{-\lambda\tau}$$

- We know the Gamma distribution is normalized by integrating the kernel:

$$1 = \int_0^\infty \frac{\lambda^{\alpha+1}}{\Gamma[\alpha+1]} \tau^\alpha e^{-\lambda\tau} d\tau$$

$$1 = \frac{\lambda^{\alpha+1}}{\Gamma[\alpha+1]} \int_0^\infty \tau^\alpha e^{-\lambda\tau} d\tau$$

$$\frac{\Gamma[\alpha+1]}{\lambda^{\alpha+1}} = \int_0^\infty \tau^\gamma e^{-\lambda\tau} d\tau$$

- Letting $\alpha = \frac{n}{2} - 1$ and $\lambda = \frac{1}{2}\left(\hat{\sigma}^2(n-k) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T \boldsymbol{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right)$

$$p[\boldsymbol{\beta} \mid \alpha, \lambda] = \frac{\Gamma[\alpha+1]}{\lambda^{\alpha+1}} = \Gamma\left[\frac{n}{2}\right] \lambda^{-\frac{n}{2}} \propto \lambda^{-\frac{n}{2}}$$

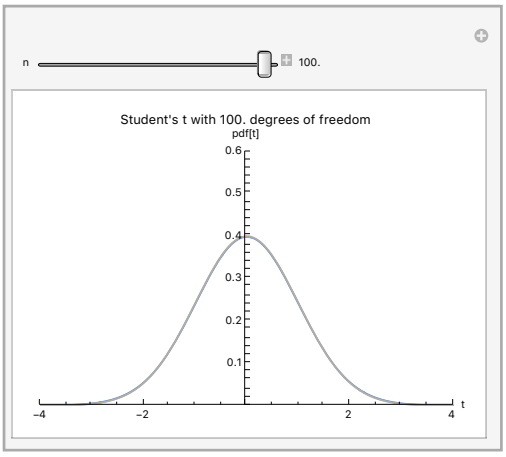
- Plugging in for $\alpha = \frac{n}{2} - 1$ and $\lambda = \frac{1}{2}\left(\hat{\sigma}^2(n-k) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T \boldsymbol{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right)$ and calling $(n-k)$ the *degrees of freedom* means that

$$p[\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}] \propto \left[(n-k) + \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^T \hat{\sigma}^2 \boldsymbol{X}^T \boldsymbol{X}\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)\right]^{-n/2}$$

- Which is the kernel of a *multivariate t-distribution.* The *Student-t distribution* for a random variable x is defined as:

$$p[x \mid \theta] = \frac{\Gamma[\theta+1]}{\Gamma\left[\frac{\theta}{2}\right]\sqrt{\pi\,\theta}} \left(1 + \frac{x^2}{\theta}\right)^{-\frac{\theta+1}{2}}$$

- Note the *t-distribution* is similar to the normal distribution, except it has heavier tails and as the sample size increases the distribution converges to a normal.



- The fact that our $\boldsymbol{\beta}$'s are *t-distributed* should come as no surprise. In regression analysis, the *t*-value is routinely calculated for $\beta_i = 0$, where it serves as an indicator of the posterior probability of a value of $\beta_i = 0$.
- Often a "*p-value*" is also calculated, which indicates the cumulative probability from the *t-distribution* of β_i being 0 or below (above, in the case where $\hat{\beta}_i < 0$).
- Conventional statistical analysis uses this *t-statistic* to "accept" or "reject" the hypothesis that $\beta = 0$.
- A small *t-value* indicates is that the data are not providing strong evidence for the hypothesis, not that the practical effect is small.
- One reason the data may not provide much evidence for an effect, even if the effect is practically important, is that there is not enough variation of x in the data.

- Conversely, small regression coefficients with large *t-values* indicate that the effect of x on y cannot be very large, unless there is something extremely peculiar about the particular data set.
- A large *t-value*, as we have seen, tells us that the posterior probability for β is tightly concentrated around $\hat{\beta}$.
- It makes more sense to look at the whole posterior probability distribution for β to get an idea of what evidence the regression with the particular data set can provide to understand the relationship between x and y.

Marginal posterior distribution of σ^2

- The marginal posterior distribution of the scalar σ^2 is somewhat less tedious as we can separate the posterior form:

$$p\left[\sigma^2 \mid \boldsymbol{X}, \boldsymbol{y}\right] = \int_{-\infty}^{\infty} p\left[\boldsymbol{X}, \boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2\right] p[\boldsymbol{\beta}]\, p\left[\sigma^2\right] d\boldsymbol{\beta}$$

$$\propto \sigma^{-n-1}\, e^{-\frac{1}{2\sigma^2}\left(\hat{\sigma}^2(n-k)\right)} \int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}\left(\left(\boldsymbol{\beta}-\hat{\beta}\right)^T \boldsymbol{X}^T \boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\beta}\right)\right)} d\beta$$

- The integral is a *k-dimensional* kernel of a multivariate normal distribution which is integrated to give the familiar normalizing factor:

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2\sigma^2}\left(\left(\boldsymbol{\beta}-\hat{\beta}\right)' \boldsymbol{X}' \boldsymbol{X}\left(\boldsymbol{\beta}-\hat{\beta}\right)\right)} d\beta = \left(2\pi\sigma^2\right)^{k/2}$$

- This implies:

$$p\left[\sigma^2 \mid \boldsymbol{X}, \boldsymbol{y}\right] \propto \sigma^{-n-1}\, e^{-\frac{1}{2\sigma^2}\left(\hat{\sigma}^2(n-k)\right)} \left(2\pi\sigma^2\right)^{k/2}$$

$$\propto \left(\sigma^2\right)^{-\frac{(n-k)}{2}-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\left(\hat{\sigma}^2(n-k)\right)}$$

- This is the kernel of the Inverse-Gamma distribution

$$\mathcal{IG}\left[\alpha = \frac{1}{2}(n-k-1),\, \beta = \frac{1}{2}\hat{\sigma}^2(n-k)\right] \propto \left(\sigma^2\right)^{-(\alpha+1)} e^{-\beta/\sigma^2}$$

- These inferential results are exactly the same that we would get from a standard likelihood analysis of the linear model. Therefore, we can see that the standard maximum likelihood solution is equivalent to a Bayesian solution with uniform priors and is therefore a special case of a Bayesian model.

Example

- As an example let's examine Kaldor's stylized facts about the importance of the manufacturing sector for overall economic growth.
- In 1949 Petrus Verdoorn published his observations on the economic statistics of a recovering post-WWII Europe. He discovered that there was a fairly constant long-term relationship between the growth of labor productivity and industrial output.
- He found there to be on average an output elasticity of productivity approximately equal to 0.45. Kaldor in 1967 argued it is the economic factors such as the dynamic relationship between supply and demand and the relative importance of the industrial sector in an economy that defines its growth potential, or conversely its limits to growth.

- Kaldor's second law (also known as Verdoorn's Law) states that there is a strong positive relation between the rate of growth of labor productivity ($\hat{\xi}$) and the growth of output ($\hat{X}$) where the "hat" represents the log derivative:

$$\hat{\xi}_t = -\beta_0 + (1 - \beta_1)\,\hat{X}_t$$

- We can decompose output into labor productivity and employment $X = \frac{X}{L} L = \xi L$. Hat calculus tells us that taking the log derivative to a variable Z is equal to the rate of growth of Z because:

$$\frac{d\,\text{Log}[Z]}{dt} = \frac{d\,\text{Log}[Z]}{dZ}\,\frac{dZ}{dt} = \frac{dZ/dt}{Z} = \frac{\dot{Z}}{Z} = \hat{Z}$$

- Applied to the growth of output, $\hat{X} = \hat{\xi} + \hat{L}$ which is the growth of labor productivity ($\hat{\xi}$) plus the growth of employment ($\hat{L}$). Therefore, we can rewrite Kaldor's second law as:

$$\hat{L}_t = \beta_0 + \beta_1\,\hat{X}_t$$

- The most important coefficient in the Kaldor-Verdoorn regression is β_1 which is called the Kaldor-Verdoorn coefficient, it gives the partial effect of output growth on employment growth.
- If the rate of growth of output increases 1 percent then the rate of growth of employment increases by β_1 percent.
- β_1 is the *returns to scale* of labor productivity such that when $(1 - \beta_1) < 1$ the reciprocal $\frac{1}{1-\beta_1}$ is greater than unity suggesting increasing returns to scale are present.

In R

Nonlinear Regression

- Linear regression tends to two types of misleading results both arising from the fact that linear regression averages out local correlations over the whole domain of the sample.
- First, linear methods tend to miss non-linear correlations which may be positive in one part of the domain and negative in another. The average correlation may be zero, even when there is a meaningful dependence.
- Second, linear regression is extremely sensitive to outliers, which can create the spurious impression of overall correlation between variables which actually vary independently.
- Linear relationships are best adapted to understanding equilibrium systems undergoing small perturbations from stable equilibrium configurations. It is the assumed small size of variations that make the linear specification plausible.
- Therefore, the most favorable context for the linear hypothesis is when the data describe "small" fluctuations of x and y around some equilibrium.
- In this setting we expect the term $y''[0]\,x$ from the Taylor expansion to be small, and so we will not make much of an error by neglecting to take it into account.
- Unfortunately economic data rarely comes from this type of local fluctuation, and often the range of economic data is quite large.
- Furthermore, economic theories of almost all types emphasize non-linearities as a key element in the theory.
- For example, neoclassical economics is built around the non-linearity of indifference curves and production isoquants, which accounts for downward-sloping demand curves and upward sloping supply curves.

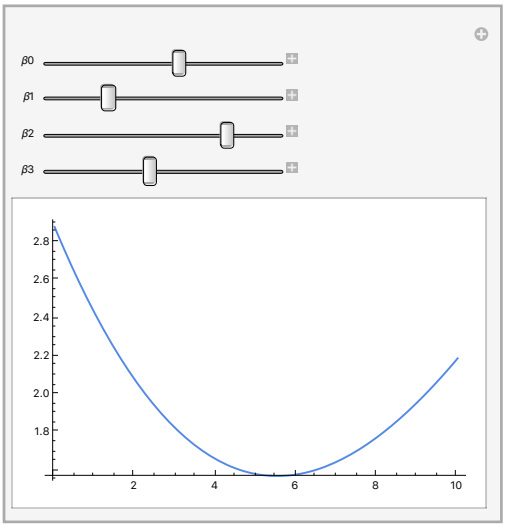
- Similarly post-Keynesian economics invokes many non-linearities, such as Minsky effects of balance sheet compositions, Goodwin labor market specifications, and the like. Marxian economics also introduces effective non-linearities in key relationships such as the rising organic composition of capital.
- Thus there are strong theoretical reasons not to rule out non-linear effects in economic data analysis. In Bayesian terms this means including a wider set of models than linear models in the prior.

Local and Global Non-Linear Specifications

- One way to do this would be to include the quadratic, and perhaps higher-order coefficients, such as cubics in the model.

$$y[x] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

- If the posterior probability given the data gives a very low posterior probability to models with β_2 or β_3 very different from zero, that is evidence for the linear effect being much larger than the quadratic or cubic effects. But at least we have not ruled out non-linear effect *a priori*.
- Although a higher-order specification is considerably more general than the linear regression hypothesis, it has global implications.
- In the cubic hypothesis, the posterior probability of β_2 still depends on the whole data set. Changing β_2 will change the fit and hence the error from which the regression posterior is derived at every data point.



- Thus it is impossible with this type of model to adjust the fit of the data for particular sub-domains of the data. The problem is that a cubic specification is a *global* specification, even though it is non-linear, since the posterior probability of the parameters depend jointly on all the data.
- This same issue arises with other common transformations of the data, such as taking logarithms or exponentials. (Both logarithms and exponentials can be approximated over any finite domain by polynomials.)
- A neoclassical constant elasticity production function specification such as:

$$y[x] = A\, x^{\eta}, \text{ or equivalently}$$

$$\text{Log}[y[x]] = \text{Log}[A] + \eta\, \text{Log}[x]$$

- Will lead to a posterior probability for the elasticity η that depends on the whole data set.

In R

Local Non-Linear Specifications

- There are (at least) two alternatives that allow for *local* modification of the deterministic model to fit particular sub-domains of the data.
- One is the Hodrick-Prescott filter specification that we will look at with time series later in the semester.
- A second widely used local technique is *loess* or local regression. The idea here is to find regression fits for subdomains of the data separately, which in principle allows the slope of the deterministic model to vary in any manner over the domain of the data.
- Loess regression is a locally weighted scatter plot smoothing process based on local polynomial fits.
- In practice the loess method overlaps the domains so as to produce a smooth deterministic model. Loess requires the analyst to choose a parameter that controls the width of the subdomains, and the resulting high-posterior probability deterministic models will depend to some degree on this parameter.

In R

Binary Choice

- The binary choice situation is very frequently encountered in economic contexts, where observed behavior often has to be "coded" in categorical (particularly yes-no, or 0-1) form.
- We might have data on a sample of households including information on household characteristics like the size of the household, the age of the head of the household, total household income, and other relevant variables, and also a qualitative variable that is 1 if the household is in debt and 0 if it is not. It could be useful for various reasons to know what impact variables like household income, family size, etc., might have on household indebtedness.
- To keep the model initially simple, suppose we include only household income as a determinant of household indebtedness. Then the data take the form of pairs $\{x_i = \text{income}_i, y_i = \text{indebted}_i\}$ where i indexes the particular household observed.
- We might try to explain household indebtedness as a linear function of household income:

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

- But this is not going to work in general, because indebtedness can only take on the values 0 and 1, while the right-hand side of this model will give a continuum of values.

In R

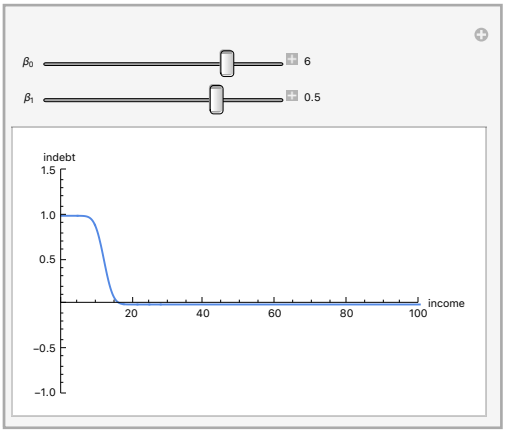
Binary Choice

- One commonly adopted approach to this problem is to shift attention to modeling the *frequency* of indebtedness conditional on different levels of income. If we let the frequency of indebtedness be $p[x]$, then the probability that $p[y = 1 \mid x]$ (household with income level x is in debt) is $p[x]$ and $p[y = 0 \mid x] = 1 - p[x]$. More compactly we can write this as a Binomial distribution:

$$p[y \mid x] = p[x]^y \, (1 - p[x])^{1-y}$$

- where $p[x]$ is a function *linking* the actual y to the estimated y in an econometric model. If $p[x]$ is just a linear function then the problem reduces to the linear model.

- We need to transform the dichotomous y into a continuous variable $y \in (-\infty, \infty)$ So we need a link function $p[x]$ that takes a dichotomous y and gives us a continuous, real-valued y'.
- We still don't want the frequency to be a linear function of income, since frequencies have to be bounded between 0 and 1. So we might pass the linear function through some kind of "S-shaped" sigmoidal function that is asymptotic to 0 and 1.
- What function $p[x]$ goes from the [0,1] interval to the real line? The two most commonly used link functions are the cumulative normal distribution Φ and the logit transform.



Probit

- The probit model uses the CDF of the Standard Normal Distribution as the link function, so the regression is of the form

$$y = \Phi[\beta_0 + \beta_1 x]$$

```
CDF[NormalDistribution[], β0 + β1 x]
```

$$\frac{1}{2}\,\text{Erfc}\left[\frac{-\beta_0 - x\,\beta_1}{\sqrt{2}}\right]$$

- The binomial choice model, like the Bernoulli trials model, does not make a deterministic prediction for each data observation, and thus the likelihood is inherently consistent with any possible pattern of observed behavior.
- As a result, the likelihood can be used directly for inference, without our having to add another layer of error probabilities into the formalism.
- For any model parameters β_0, β_1, and n data points $\{\{x_1, y_1\}, \ldots, \{x_n, y_n\}\}$ the likelihood for this model is the product:

$$p[\{\{x_1, y_1\}, \ldots, \{x_n, y_n\}\} \mid p[x] = \beta_0 + \beta_1 x] = \prod_{i=1}^{n} p[x_i]^{y_i} (1 - p[x_i])^{1-y_i}$$

- One advantage of this model is that it does not rule out any particular household from any combination of income and indebtedness status.

In R

Logit

- A alternative to the probit model is to specify the link function as the logistic function $p[x] = \frac{e^x}{1+e^x}$ based off of the log odds ratio

$$\text{Log}\left[\frac{p}{1-p}\right] = x$$

$$\frac{p}{1-p} = e^{x}$$

$$p = \frac{e^{x}}{1+e^{x}}$$

- For the linear predictive model this becomes

$$\text{Log}\left[\frac{p[y = 1 \mid x, \beta]}{1 - p[y = 1 \mid x, \beta]}\right] = \beta_0 + \beta_1 x$$

$$\frac{p[y = 1 \mid x, \beta]}{1 - p[y = 1 \mid x, \beta]} = e^{\beta_0 + \beta_1 x}$$

$$p[y = 1 \mid x, \beta] = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Range of Regression

- The standard distributions used in regression analysis have natural derivations from simple probability models.
- The Binomial Distribution is motivated from counting exchangeable outcomes.
- The Normal Distribution applies to a random variable that is the sum of many exchangeable or independent terms.
- The Poisson and Exponential Distributions arise as the number of counts and the waiting times, respectively, for events modeled as occurring exchangeably in all time intervals; i.e., independently in time, with a constant rate of occurrence.

Poisson Regression

- The Poisson distribution can be used to model unbounded count data $x \in \{0, 1, 2, \ldots\} \sim \text{Poisson}[\lambda]$ with the likelihood

$$p[x] = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda x_i} = \frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!} e^{-n\lambda} \propto \lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}$$

- The Poisson distribution has a single parameter λ, which is the mean of the distribution and also the variance.
- In the Poisson model we have a sample of n observations $y_1, y_2, \ldots, y_n$ which can be treated as realizations of independent Poisson random variables, with $Y_i \sim \text{Poisson}[\lambda_i]$ for $\lambda > 0$.
- Letting the mean (and variance) λ depend on a vector of explanatory variables x_i we have $\lambda[x]$. We could assume that the mean count λ is a linear function of x such that

$$\lambda_i[x] = \beta_0 + \beta_1 x_i$$

- However, in this model the linear predictor on the right hand side can assume any real value, whereas the Poisson mean has to be non-negative.
- Instead we can model the logarithm of the mean using a linear model. Taking $\mu = \text{Log}[\lambda]$ the transformed log-linear model is

$\mu_i = \text{Log}[\lambda_i[x]] = \beta_0 + \beta_1 x_i$

- In this case the regression coefficient β_1 represents the expected change in the log of the mean count per unit change in the predictor x_i. Increasing x_i by one unit is associated with an increase of β_1 in the log of the mean count.
- However, for estimation we need to transform the model parameter so that it is not expressed logarithmically.

$\lambda_i[x] = e^{\beta_0+\beta_1 x_i}$

In R

Quantile Regression

- Regression models typically refer to the conditional distribution of y given x, parameterized by θ as $p[y \mid x, \theta]$.
- The quantile regression approach models the conditional quantile of y given x:

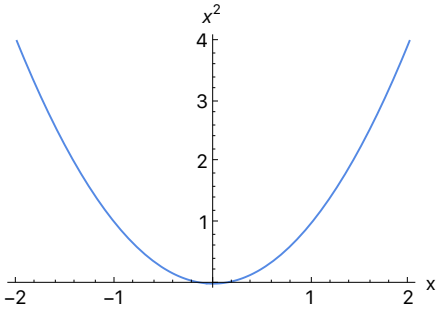
$Q_\tau[y \mid x, \theta] \text{ with } 0 < \tau < 1$

- For example, a regression on the median of y would have $\tau = 0.5$.
- If $p[y \mid x]$ is symmetric, then the median regression is identical to the mean regression, otherwise, they are different. In quantile regression we have:

$$E[y \mid x] = \int_{\tau=0}^{1} Q_\tau[y \mid x, \theta]\, d\tau$$

- If x and y are distributed as a Bivariate Normal Distribution then symmetry of $p[y \mid x]$, all quantile curves will be parallel.
- Otherwise we may have different relationships across the distribution $p[y, x]$.
- Presented with a random sample $\{y_1, \ldots, y_n\}$ we define the sample ***mean*** as the solution to the problem of minimizing a sum of squared residuals.

$\min_{\mu \in \mathbb{R}} \sum_i (y_i - \mu)^2$




```
y = RandomInteger[100, 100]
```

```
{21, 76, 71, 92, 64, 90, 28, 21, 14, 75, 3, 13, 40, 91, 11, 70, 15, 58, 14, 2, 39, 51, 36, 28,
 20, 97, 19, 59, 12, 100, 65, 72, 21, 14, 87, 80, 44, 7, 48, 55, 11, 87, 68, 87, 47, 37, 97,
 36, 22, 60, 85, 63, 33, 25, 36, 93, 27, 50, 88, 59, 36, 10, 13, 54, 100, 63, 44, 23, 55, 72,
 11, 55, 74, 11, 0, 78, 95, 41, 80, 29, 86, 42, 13, 23, 40, 69, 52, 0, 37, 66, 18, 100, 85,
 25, 35, 83, 36, 29, 99, 19}
```

```
FindMinimum[Total@(y - μ)^2, μ]
```

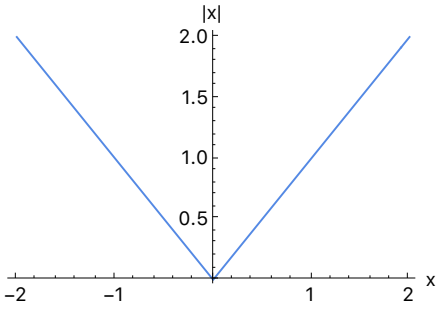
```
{86076.8, {μ → 48.35}}
```

```
Mean[y] // N
```

```
48.35
```

- We can define the ***median*** as the solution to the problem of minimizing a sum of absolute residuals.

$$\min_{\mu \in \mathbb{R}} \sum_i \left| y_i - \mu \right|$$




```
FindMinimum[Total@Abs[y - μ], μ] // Quiet
```

```
{2555., {μ → 44.}}
```

```
Median[y] // N
```

```
44.
```

- Replace the scalar μ by a parametric function $\mu[x, \beta]$ and we get an estimate of the conditional expectation function $E[y \mid x, \beta]$:

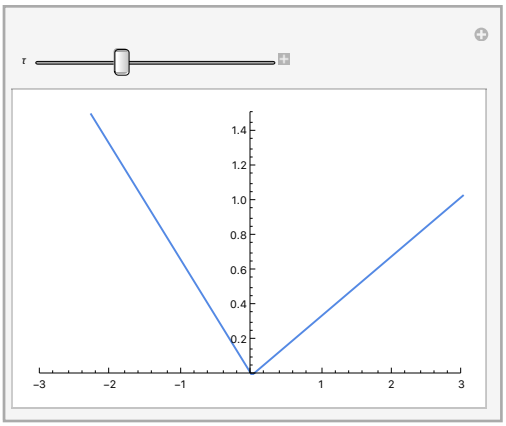
$$\min_{\mu \in \mathbb{R}} \sum_i (y_i - \mu[x_i, \beta])^2$$

- In fact we can define any quantiles as a similar optimization problem:

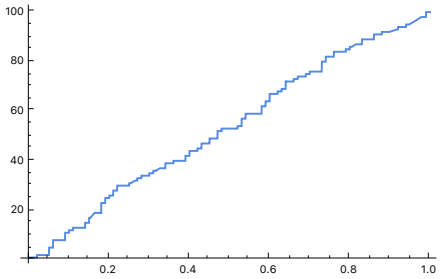
$$\min_{\xi \in \mathbb{R}} \sum \rho_\tau[y_i - \xi]$$

- In this case ρ is a loss function that weights the data around a quantile τ.

$$\rho_\tau[y] = y(\tau - I[y < 0]) = \begin{cases} y\,\tau & \text{if } y < 0 \\ y(\tau - 1) & \text{if } y > 0 \end{cases}$$



```
Plot[Quantile[y, {x}], {x, 0, 1}]
```



- In quantile regression we follow the same logic as above. To get an estimate of the conditional median replace the scalar ξ by the parametric function $\xi[x_i, \beta]$ and set $\tau = 1/2$.
- To obtain estimates of the other conditional quantile functions we simply replace absolute values by ρ_τ and solve

  $\min_{\xi \in \mathbb{R}} \sum \rho_\tau[y_i - \xi[x_i, \beta]]$

In R

- Let's examine Ernst Engel's data on food consumption and income.
- The dispersion of food expenditure to increase along with its level as household income increases.
- The spacing of the quantile regression lines also reveals that the conditional distribution of food expenditure is skewed to the left.
- We can see this from the narrower spacing of the upper quantiles indicating high density and a short upper tail and the wider spacing of the lower quantiles indicating a lower density and longer lower tail.
- Also, the conditional median is higher than the conditional mean.