ECON 557 – Advanced Data Analysis

Michael T. Sandfort

Department of Economics
Masters in Applied Economics Program
Georgetown University

January 27, 2023

Maximum Likelihood (ML): The Likelihood Function

- If the random variable $y|\mathbf{x}$ is drawn from the population $f(y|\mathbf{x}, \boldsymbol{\theta})$ with parameter $\boldsymbol{\theta}$, then we can easily compute any feature of this conditional distribution, including the conditional expectation function, from knowledge of $\boldsymbol{\theta}$ and $f(y|\mathbf{x}, \boldsymbol{\theta})$.

- What's more, the probability (or likelihood) of observing any given $y|\mathbf{x}$ is just $f(y|\mathbf{x}, \boldsymbol{\theta})$, so let

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}, y|\mathbf{x}) = f(y|\mathbf{x}, \boldsymbol{\theta})$$

  This is called the likelihood function for random variable $y|\mathbf{x}$. But it is a function of the parameter $\boldsymbol{\theta}$, not of $y|\mathbf{x}$, and it is not normalized (so not a distribution).

Maximum Likelihood (ML): The Likelihood Function

- We often work with the log-likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) \equiv \ln L(\boldsymbol{\theta}) = \ln f(y|\mathbf{x}, \boldsymbol{\theta})$$

  rather than the likelihood function.

- This is legitimate because we are about to try to maximize the likelihood.

- Any monotone transformation of an optimization problem preserves the optimizer, although it does not typically preserve the optimum.

Maximum Likelihood (ML): Conditioning

- Strictly speaking, the probability of observing a given draw $(y, \mathbf{x})$ given parameter $\boldsymbol{\theta}$ is

$$f_{y|\mathbf{x}}(y|\mathbf{x}, \boldsymbol{\theta}) f_{\mathbf{x}}(\mathbf{x}|\boldsymbol{\theta})$$

  where $f_{\mathbf{x}}(\cdot)$ is the marginal density of $\mathbf{x}$.

- Our likelihood function from the previous slide is actually a conditional likelihood function.

- So long as $f_{y|\mathbf{x}}$ and $f_{\mathbf{x}}$ depend on mutually exclusive sets of parameters, maximizing L over the relevant parameters for $f_{y|\mathbf{x}}$ does not depend on $f_{\mathbf{x}}$, so estimates of $\boldsymbol{\theta}$ will be consistent.

- In this case, both the marginal distribution $f_{\mathbf{x}}$ and references to "conditional" likelihood are usually dropped.

Maximum Likelihood (ML): The Optimal Parameter

- With y a random variable, the $\boldsymbol{\theta}$ which maximizes $\mathcal{L}(\boldsymbol{\theta}) = \ln f(y|\mathbf{x}, \boldsymbol{\theta})$ is also a random variable (same for $L(\boldsymbol{\theta})$).
- This makes it unappealing as a "population feature" for us to target.
- Instead, we take as our target in the population the value $\boldsymbol{\theta}_o$ which maximizes the expected likelihood

$$\boldsymbol{\theta}_o \equiv \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} E[\mathcal{L}(\boldsymbol{\theta})]$$

  where the expectation is taken with respect to $f(y|\mathbf{x}, \boldsymbol{\theta})$.
- We want a $\boldsymbol{\theta}$ which gets $f(y|\mathbf{x}, \boldsymbol{\theta})$ close to the true $f(y|\mathbf{x}, \boldsymbol{\theta}_o)$, across all realizations $(y, \mathbf{x})$.

A Loss Function for ML

- One measure of the difference between the two densities $f(y|\mathbf{x}, \boldsymbol{\theta})$ and $f(y|\mathbf{x}, \boldsymbol{\theta}_o)$ for some draw $(y, \mathbf{x})$ is

$$\ln f(y|\mathbf{x}, \boldsymbol{\theta}_o) - \ln f(y|\mathbf{x}, \boldsymbol{\theta}) = \ln \left[\frac{f(y|\mathbf{x}, \boldsymbol{\theta}_o)}{f(y|\mathbf{x}, \boldsymbol{\theta})} \right]$$

- While this can be either positive or negative for a given draw $(y, \mathbf{x})$, Jensen's Inequality implies that

$$\begin{aligned} KL(f_{\boldsymbol{\theta}_o}, f_{\boldsymbol{\theta}}) &\equiv E_{\boldsymbol{\theta}_o}[\ln f(y|\mathbf{x}, \boldsymbol{\theta}_o) - \ln f(y|\mathbf{x}, \boldsymbol{\theta})] \\ &= \int \ln \left[\frac{f(y|\mathbf{x}, \boldsymbol{\theta}_o)}{f(y|\mathbf{x}, \boldsymbol{\theta})} \right] f(y|\mathbf{x}, \boldsymbol{\theta}_o) dy \\ &= - \int \ln \left[\frac{f(y|\mathbf{x}, \boldsymbol{\theta})}{f(y|\mathbf{x}, \boldsymbol{\theta}_o)} \right] f(y|\mathbf{x}, \boldsymbol{\theta}_o) dy \\ &\geq \ln \left[\int \frac{f(y|\mathbf{x}, \boldsymbol{\theta})}{f(y|\mathbf{x}, \boldsymbol{\theta}_o)} f(y|\mathbf{x}, \boldsymbol{\theta}_o) dy \right] \\ &= \ln(1) = 0 \end{aligned}$$

A Loss Function for ML

- $KL(f_{\boldsymbol{\theta}_o}, f_{\boldsymbol{\theta}})$ is known as the Kullback-Leibler divergence between $f_{\boldsymbol{\theta}}$ and $f_{\boldsymbol{\theta}_o}$.
- Viewed as a loss function for rating hypotheses $f_{\boldsymbol{\theta}}$, it is known as the Kullback-Leibler information criterion, or KLIC.
- KLIC has some of the properties of a distance measure:
  - The KLIC from any distribution to itself is zero, so $KL(g, g) = 0$.
- But:
  - It is not symmetric so $K(g, h) \neq K(h, g)$.
  - It does not satisfy a triangle inequality.

Risk for ML

- Following our earlier approach, we can now set up the risk function by taking $R(f_{\boldsymbol{\theta}}) = KL(f_{\boldsymbol{\theta}_o}, f_{\boldsymbol{\theta}})$ and find the Bayes-optimal hypothesis via

$$\arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} KL(f_{\boldsymbol{\theta}_o}, f_{\boldsymbol{\theta}})$$

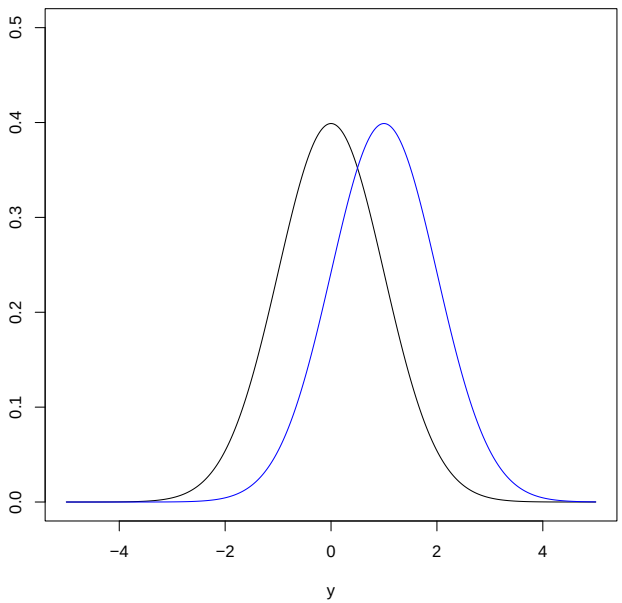
- Of course, the Bayes-optimal hypothesis is $\boldsymbol{\theta}_o$, since $KL \geq 0$ and $KL(f_{\boldsymbol{\theta}_o}, f_{\boldsymbol{\theta}_o}) = 0$.
- Why did we bother with this extra step?
- Two reasons:
  - It situates ML within the framework of risk minimization (i.e., KL loss is one variety of loss).
  - If the model is wrong, the true F may not be characterized by any $\boldsymbol{\theta}_o \in \boldsymbol{\Theta}$.

Example

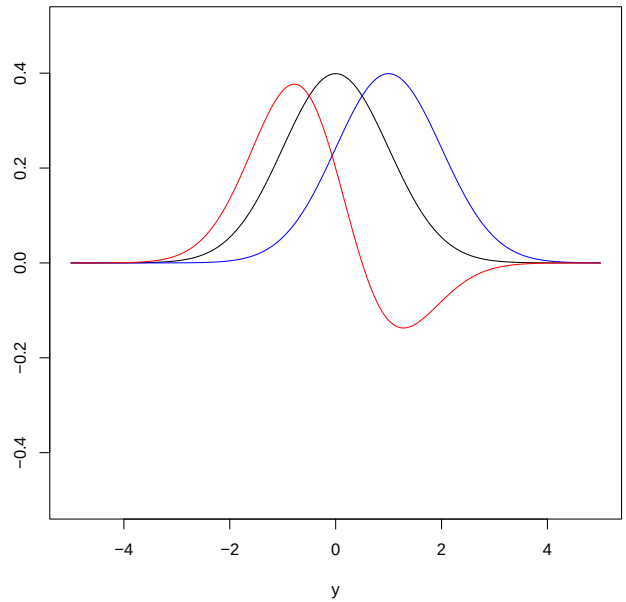
- Suppose $f(y|\mathbf{x}, \boldsymbol{\theta}) = f(y|\mathbf{x}, \boldsymbol{\theta}_o)$ for some $\boldsymbol{\theta}_o \in \Theta$.
- Then the KL loss can be minimized at zero.
- $\boldsymbol{\theta}_o$ is our Bayes-optimal hypothesis.

KL Divergence of a $\mathcal{N}(1,1)$ from a true $\mathcal{N}(0,1)$

Distributions



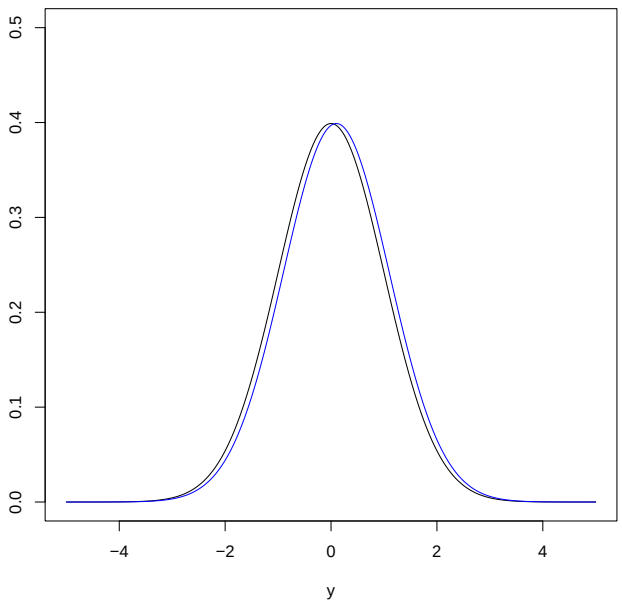
Weighted Log Ratio of Distributions



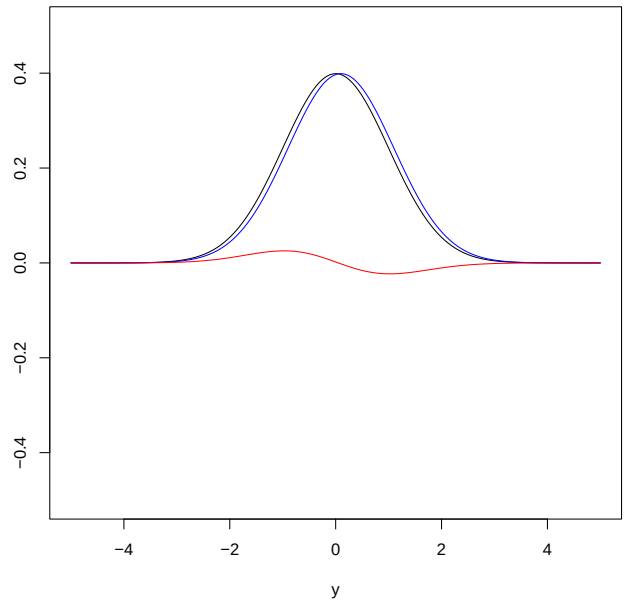
KL Divergence: 0.5.

KL Divergence of a $\mathcal{N}(0.1, 1)$ from a true $\mathcal{N}(0, 1)$

Distributions



Weighted Log Ratio of Distributions



KL Divergence: 0.005.

Example

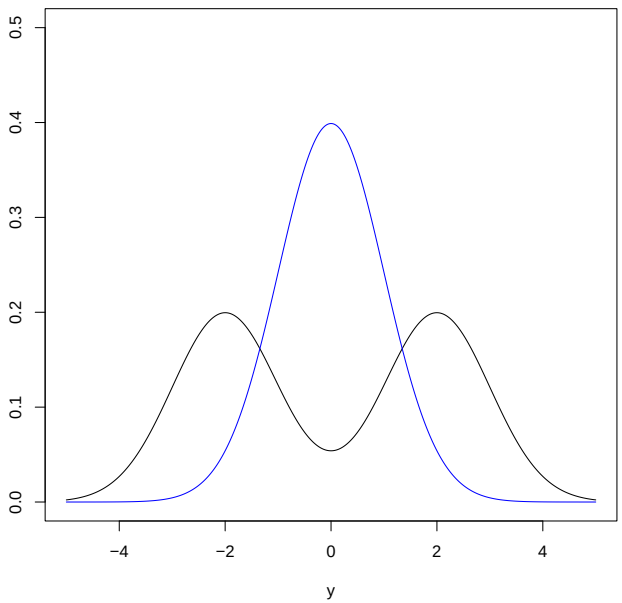
- Suppose there exists no $\boldsymbol{\theta}_o \in \boldsymbol{\Theta}$ with $f(y|\mathbf{x}, \boldsymbol{\theta}) = g(y|\mathbf{x})$ (i.e., g does not lie within the family of distributions defined by f and $\boldsymbol{\Theta}$).
- Then (with some regularity conditions) a Bayes-optimal hypothesis $\boldsymbol{\theta}_{\text{QML}}$, sometimes called the pseudo-true value, still exists.
- The pseudo-true value solves

$$\begin{aligned}
\boldsymbol{\theta}_{\text{QML}} &\equiv \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} R(\boldsymbol{\theta}) \\
&= \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_g \left[\log g(y|\mathbf{x}) - \log f(y|\mathbf{x}, \boldsymbol{\theta}) \right] \\
&= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathbb{E}_g \left[\log f(y|\mathbf{x}, \boldsymbol{\theta}) \right]
\end{aligned}$$

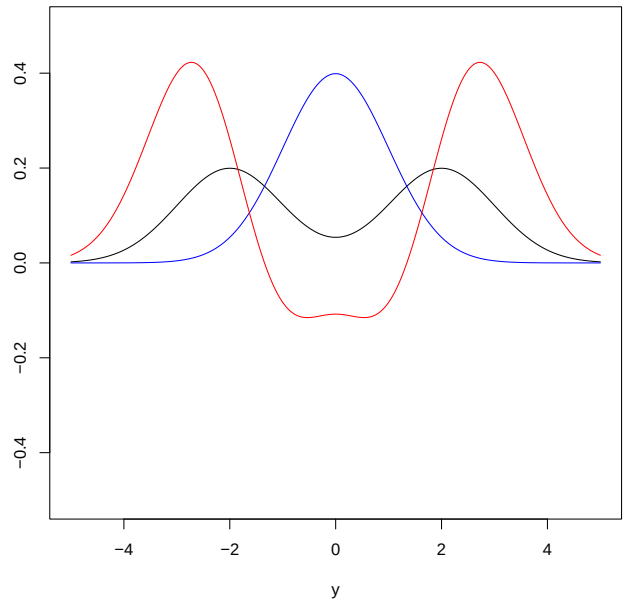
  since $g(y|\mathbf{x})$ is not a function of $\boldsymbol{\theta}$ and maximizing $-f$ is the same as minimizing f.
- This is nothing more than the mis-specified maximum likelihood problem using the family $f_{\boldsymbol{\theta}}$ to fit g.
- Whether $\boldsymbol{\theta}_{\text{QML}}$ captures meaningful information about g is a separate question.

KL Divergence of a $\mathcal{N}(0,1)$ from a true mixed $\frac{1}{2}\mathcal{N}(-2,1) + \frac{1}{2}\mathcal{N}(2,1)$

Distributions



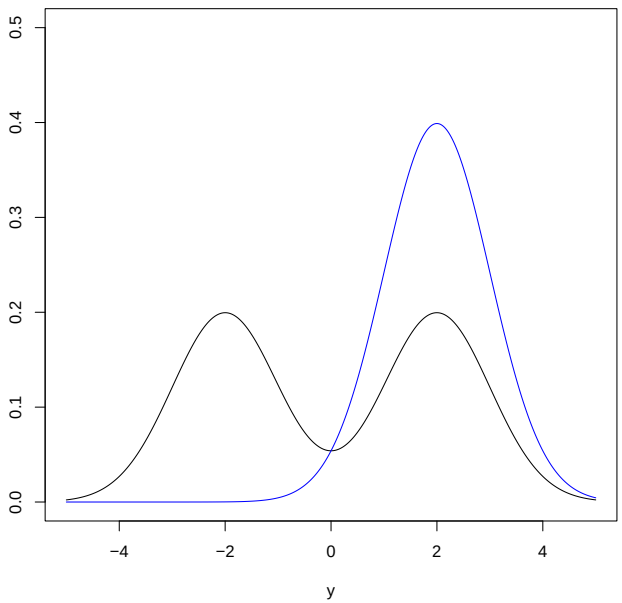
Weighted Log Ratio of Distributions



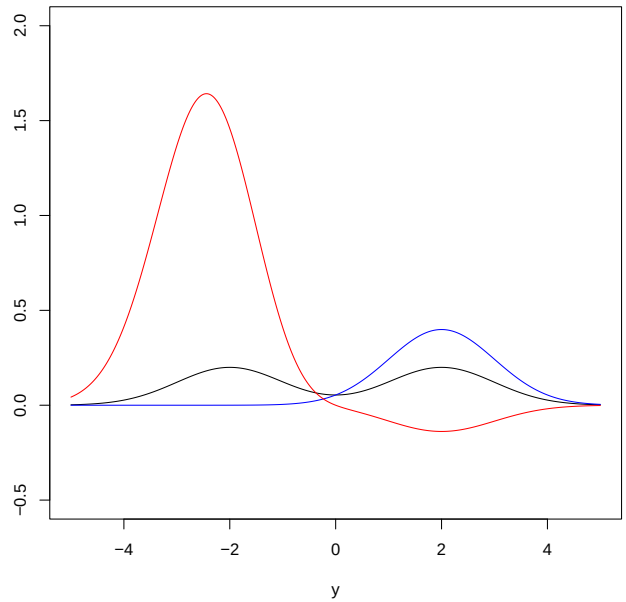
KL Divergence: 1.357.

KL Divergence of a $\mathcal{N}(2,1)$ from a true mixed $\frac{1}{2}\mathcal{N}(-2,1)+\frac{1}{2}\mathcal{N}(2,1)$

Distributions



Weighted Log Ratio of Distributions



KL Divergence: 3.354.

The Return of the Exponential Diffusion Family

- Suppose that the family of f_θ is a one-parameter EDF. Then by what we have already shown,

$$\begin{aligned}
\theta_{\text{QML}} &= \arg\max_{\theta\in\Theta} \mathbb{E}_g \left[\log f(y|\mathbf{x},\theta)\right] \\
&= \arg\max_{\theta\in\Theta} \mathbb{E}_g \left[\log \exp\left\{\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right\}\right] \\
&= \arg\max_{\theta\in\Theta} \mathbb{E}_g \left[\frac{y\theta - b(\theta)}{\phi} + c(y,\phi)\right] \\
&= \arg\max_{\theta\in\Theta} \mathbb{E}_g \left[y\theta - b(\theta)\right]
\end{aligned}$$

  from which a differential characterization gives (as before) that $\mathbb{E}_g[y] = b'(\theta)$.

- Since $b(\cdot)$ is a known function, θ_{QML} reveals the mean of y.

Generalized Linear Modeling

- If you were watching closely, you probably noticed that $\mathbf{x}$ disappeared from the previous page.
- We are about to put it back – by relating it to the natural parameter θ.
- As in your past work, consider a linear model for the covariates $\mathbf{x}$ of the form

$$\eta = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m = \mathbf{x}\boldsymbol{\beta}$$

- The value η is usually called an index value.
- Next, we allow the mean on y to be related to the index through an invertible response function, as

$$\mathbb{E}[y|\mathbf{x}, \boldsymbol{\beta}] = h(\eta) = h(\mathbf{x}\boldsymbol{\beta})$$

- This is what we do, for example, when we let $\mathbb{P}[y = 1|\mathbf{x}, \boldsymbol{\beta}] = \Phi(\mathbf{x}\boldsymbol{\beta})$ in a probit model.

Bringing It All Together

- For y in a one-parameter EDF, we now have two expressions involving the mean on y:

$$\mu = \mathbb{E}[y|\theta, \phi] = b'(\theta)$$
$$\mu = \mathbb{E}[y|\boldsymbol{x}, \boldsymbol{\beta}] = h(\mathbf{x}\boldsymbol{\beta}) = h(\eta)$$

- From this, it follows that

$$\theta = (b')^{-1}(\mu) = (b')^{-1}(h(\mathbf{x}\boldsymbol{\beta}))$$

- Keeping in mind that the form of $b(\cdot)$ is dictated by the distribution of outcomes y, notation is vastly simplified if we take $h = b'$. If so, our linear index is exactly the natural parameter, or $\theta = \eta$.
- The particular $h = b'$ is called the natural (or canonical) response. Its inverse g is called the natural (or canonical) link function.
- While taking h as the natural response is highly tractable, it may not be right. Fortunately, other choices of response function (or equivalently, link function) are not difficult to implement.

Example: The Normal Distribution

- We showed last time that the normal distribution has $\mu = b'(\theta) = \theta$.
- So the natural response function $h(\eta) = \eta$ is just the identity function.
- Since the identity function is its own inverse, the natural link function g is also the identity.
- Classical normal regression has a linear CEF with $E[y|\mathbf{x}] = \mu = h(\eta) = h(\mathbf{x}\boldsymbol{\beta}) = \mathbf{x}\boldsymbol{\beta}$, so this should be familiar.

Example: The Bernoulli Distribution

- We showed last time that the Bernoulli distribution has $\mu = b'(\theta) = \frac{e^\theta}{1+e^\theta}$.
- So the natural response function is $h(\eta) = \frac{e^\eta}{1+e^\eta}$.
- The natural link function g is given by inverting $h(\eta) = \mu$, so $g(\mu) = \log \frac{\mu}{1-\mu}$.
- In logit regression we take $\mathbb{P}[y = 1|\mathbf{x}] = E[y|\mathbf{x}] = \mu = h(\eta) = h(\mathbf{x}\boldsymbol{\beta})$ with h the logistic distribution $\Lambda(s) = \frac{e^s}{1+e^s}$.
- Probit regression is also straightforward – set $h(\eta) = \Phi(\eta)$ (the standard normal distribution). Evidently, this response function is not the natural response function for the Bernoulli.

Example: The Poisson Distribution

- We showed last time that the normal distribution has $\mu = b'(\theta) = e^{\theta}$.
- So the natural response function is $h(\eta) = e^{\eta}$.
- The natural link function g inverts h, so $g(\mu) = \log \mu$.
- Remember that Poisson regression is common for count data.

Estimation: Definitions

Estimation requires the following vocabulary:

- A population, described by its distribution function F and which we do not directly observe.[1]
- A feature of the population, $\phi(F)$ in which we are interested.
- A sample $D_N \equiv \{(\mathbf{x}_i, y_i)\}_{i=1}^N$ of data (also known as the training sample) drawn from the distribution F.
- A statistic, T, which is just a function of the sample: $T(\{(x_i, y_i)\}_{i=1\ldots N})$.

[1] For brevity in notation, we now drop the explicit notation $F_{\mathbf{x},y}$.

Population Features

Examples of univariate and bivariate populations were provided earlier.
Examples of population features include:

- Mean and other raw moments.
- Oracle predictors (mean, median, quantiles, PLP, CMF, Bayes classifier)
- Dispersion (variance, kurtosis) and other central moments.
- Correlations and covariances.
- Distribution F itself, or its density f, possibly via a parametric specification $F_{\boldsymbol{\theta}}$.

From Statistics to Estimation

An estimator of the population feature $\phi(F)$ is nothing more than a statistic T which we point at $\phi(F)$. If we choose our estimator well, it will have good properties:

- Unbiased
- Consistent
- Asymptotically Normal
- Efficient

If we choose badly, then it need have none of those properties.

Estimation: Parametric Estimation

One approach to estimation is parametric estimation. In this approach, we take F to be known up to a collection of parameters $\boldsymbol{\theta}$ and write $F_{\boldsymbol{\theta}}$ for the parametric distribution.

- Start by taking $\phi(F_{\boldsymbol{\theta}}) \equiv \boldsymbol{\theta}$ (naturally!). Often, the parameters $\boldsymbol{\theta}$ are simply the population mean, variance, etc. as for the normal or exponential distributions.
- Define an estimator $\widehat{\boldsymbol{\theta}} \equiv T(\cdot)$.
- Any other population features of interest γ can be computed by taking $\widehat{\gamma} = \gamma(F_{\widehat{\boldsymbol{\theta}}})$.

Is the OLS estimator, as you have learned and used it, a parametric estimator?

Estimation: Example

In an unknown population F, our feature of interest is the population mean μ. How to estimate from $\{y_i\}_{i=1\ldots N}$?

- One possibility is the sample mean

$$T_1(\{y_i\}) \equiv \overline{y}_N \equiv \frac{1}{N}\sum_{1}^{N} y_i$$

  which uses information from all observations in the sample, weighting them equally.

- Another would be the first observation in the sample,

$$T_2(\{y_i\}) \equiv y_1.$$

- Another would be the number 42,

$$T_3(\{y_i\}) \equiv 42.$$

What are the competing merits of these three estimators?

Estimation: The Analogy Principle

- One common approach to good estimator design is called the analogy principle.
- To estimate $\phi(F)$ in the population, compute $\phi(\widehat{F}_N)$ from the sample, where $\widehat{F}_N$ is the empirical distribution.
- The analogy estimator is sometimes called the plug-in estimator.
- For population features which are built from expectations, this approach often amounts to nothing more than replacing instances of $\mathbb{E}[z]$ with $\frac{1}{N}\sum_{i=1}^{N} z_i$.

Estimation: An Example of the Analogy Principle

Suppose we want to estimate the population mean in a population F. The population feature is

$$\phi(F) = \mathbb{E}_F[y] \equiv \int y dF(y) \equiv \mu$$

By analogy, replace F with $\widehat{F}_N$ everywhere it appears above:

$$\begin{aligned}
\widehat{\mu} &= \int y d\widehat{F}_N(y) \\
&= \sum_{i=1}^{N} y_i [\widehat{F}_N^+(y_i) - \widehat{F}_N^-(y_i)] \qquad \text{fancy math!} \\
&= \sum_{i=1}^{N} y_i \left[\frac{1}{N}\right] = \frac{1}{N} \sum_{i=1}^{N} y_i \equiv \overline{y}_N
\end{aligned}$$

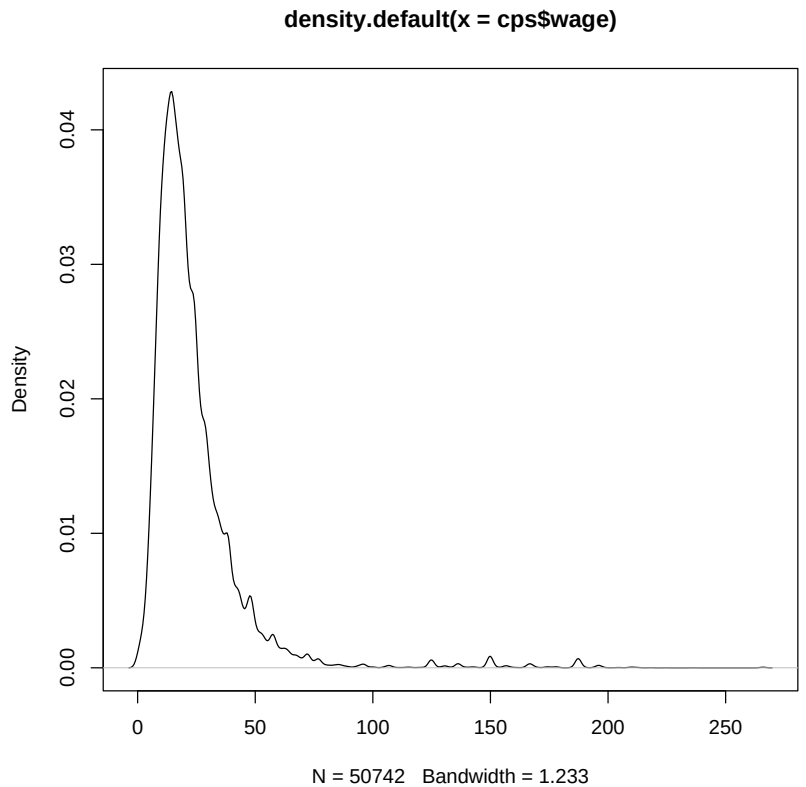
So the analogy estimator of the population mean is just the sample mean.

Estimation: Estimating the Mean Wage

```
> library(haven)
> cps <- read_dta(paste0(myDataPath,"cps09mar.dta"))
> names(cps)
 [1] "age"       "female"    "hisp"      "education" "earnings"  "hours"
 [7] "week"      "union"     "uncov"     "region"    "race"      "marital"
> N <- nrow(cps)
> cps$wage <- cps$earnings/(cps$hours*cps$week)
> mean(cps$wage)
[1] 23.90266
> (1/N)*sum(cps$wage)
[1] 23.90266
> median(cps$wage)
[1] 19.23077
> summary(cps$wage)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
  0.00038  12.82051  19.23077  23.90266  28.84615 266.05573
```

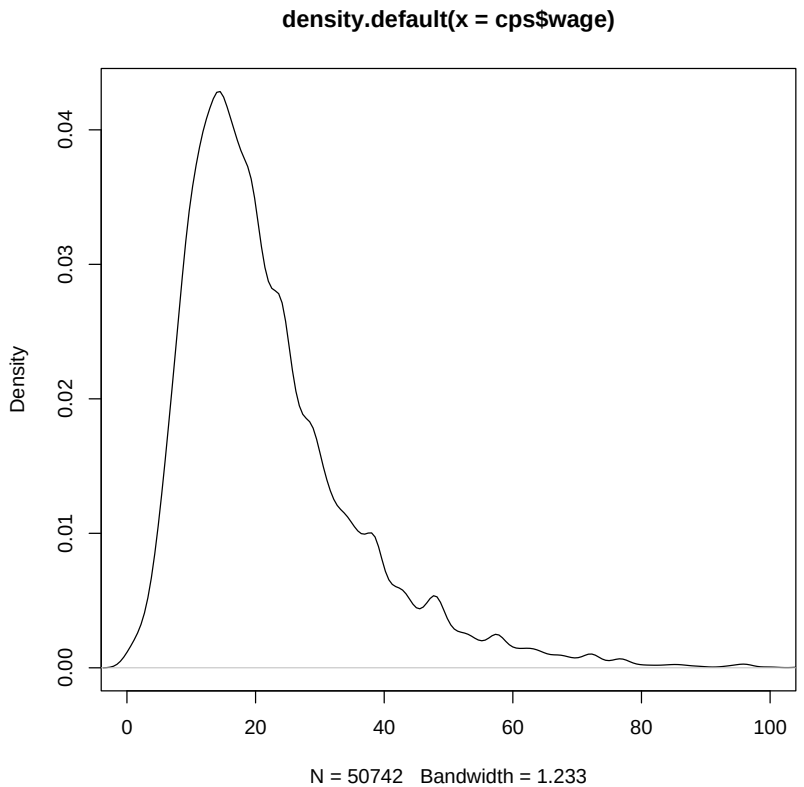
Estimation: Wage Distribution is Asymmetric

```
> # Plot of density of wage (base graphics)
> # Whole support represented
> plot(density(cps$wage))
```



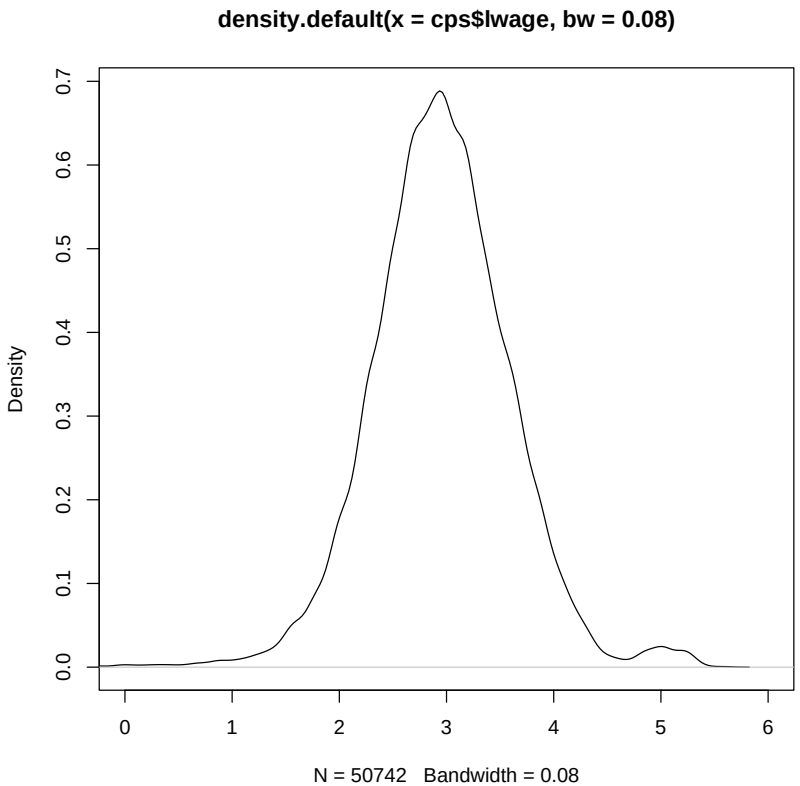
Estimation: Wage Distribution is Asymmetric

```
> # Plot of density of wage (base graphics)
> # What Hansen (p. 13) shows
> plot(density(cps$wage),xlim=c(0,100))
```



Estimation: Estimating the Mean Log Wage

```
> cps$lwage <- log(cps$wage)
> summary(cps$lwage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 -7.863   2.551   2.957   2.946   3.362   5.584
> # Clip the support like Hansen (p. 14) does
> plot(density(cps$lwage,bw=.08),xlim=c(0,6)) # bw= modifies default bandwidth
```



Estimation: Precision of the Estimates

As we have seen, the sample mean doesn't give a very good idea of the "middle" of the distribution of wages. Sample mean of log wages is more informative (if you are interested in log wages).

Even so, the sampling variances of both sample means are quite small:

```
> (1/(N-1))*sum( (cps$wage-mean(cps$wage))^2 )/N
[1] 0.008453516
> (1/(N-1))*sum( (cps$lwage-mean(cps$lwage))^2 )/N
[1] 9.002937e-06
```

Bottom line: You can get very precise estimates of parameters which are wholly uninteresting.

Estimating Population Features with Covariates

Now we'll switch gears to estimating population features involving observed covariates $\mathbf{x}$.

The roadmap for the rest of the lecture is:

- Show that OLS is the analogy estimator for the population linear predictor (LP).
- Derive an analogy estimator for the CEF, which works under special circumstances.
- Use a sample from the Current Population Survey (CPS) to demonstrate the use of both OLS and the CEF-analogy in examining the conditional distribution of log wage on
  - Sex
  - Sex and Race
- Derive an analogy estimator for the risk, which will be our key workhorse going forward. We will also show that
  - Minimum risk under MSE loss yields the OLS estimator.
  - Minimum risk under KL loss, a normal F_θ and natural response gives the OLS estimator.
  - Minimum risk under KL loss, a Bernoulli F_θ and natural response gives the logit estimator.

OLS as an Analogy Estimator

Suppose the population feature we would like to estimate is

$$\phi(F) = \boldsymbol{\beta} \equiv \left(\mathbb{E}_F[\mathbf{x}'\mathbf{x}]\right)^{-1} \mathbb{E}_F[\mathbf{x}'y].$$

Because $\boldsymbol{\beta}$ can be written as a function of expectations, the analogy principle says to replace those population expectations $\mathbb{E}_F[\cdot]$ with sample averages, getting

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= \left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i'y_i\right) \\ &= \left(\sum_{i=1}^{N}\mathbf{x}_i'\mathbf{x}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{x}_i'y_i\right) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}\end{aligned}$$

which (plus algebra) is a set of conditions $j = 1, \ldots, m$, each of the form

$$\sum_{i=1}^{N} x_{i1}x_{ij}\widehat{\beta}_1 + \cdots + x_{im}x_{ij}\widehat{\beta}_m = \sum_{i=1}^{N} x_{ij}y_i$$

or

$$\sum_{i=1}^{N} x_{ij}(y_i - x_{i1}\widehat{\beta}_1 - \cdots - x_{im}\widehat{\beta}_m) = \sum_{i=1}^{N} x_{ij}(y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}) = 0$$

OLS as an Analogy Estimator

Recalling that the sample mean of $\mathbf{y}$ can be computed as the regression of $\mathbf{y}$ on the constant vector $\boldsymbol{\iota}$, we have for the wage

```
> summary(lm(wage~1,data=cps))

Call:
lm(formula = wage ~ 1, data = cps)

Residuals:
    Min      1Q  Median      3Q     Max
-23.902 -11.082  -4.672   4.943 242.153

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.90266    0.09194     260   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.71 on 50741 degrees of freedom
```

OLS as an Analogy Estimator

And for the log wage,

```
> summary(lm(lwage~1,data=cps))

Call:
lm(formula = lwage ~ 1, data = cps)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8095  -0.3951   0.0103   0.4158   2.6375

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.946      0.003   981.9   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6759 on 50741 degrees of freedom
```

In each regression, the coefficient on the constant returns the sample mean computed earlier.

Note that the reported standard error is the square root of the previously computed sampling variance for the sample mean (we'll discuss this more next lecture).

An Analogy Estimator for the CEF

Finally, suppose the population feature $\phi(F)$ is the population best predictor (CEF), $\mathbb{E}_F[y|\mathbf{x}]$.

Let $D(y|\mathbf{x})$ represent the CDF of y conditional on $\mathbf{x}$, so that

$$\phi(F) = \mathbb{E}_F[y|\mathbf{x}] \equiv \int_y y dD(y|\mathbf{x}) \equiv \mu_{y|\mathbf{x}}$$

An Analogy Estimator for the CEF

When $\mathbf{x}$ contains only discrete values, we can define the empirical conditional distribution function

$$\begin{aligned}\widehat{D}(y|\mathbf{x}) &= \frac{\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{y_i \leq y\}\mathbf{1}\{x_{i1}=x_1\}\cdots\mathbf{1}\{x_{ik}=x_k\}}{\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{x_{i1}=x_1\}\cdots\mathbf{1}\{x_{ik}=x_k\}} \\ &= \frac{1}{N_{\mathbf{x}}}\sum_{i\in N_{\mathbf{x}}}\mathbf{1}\{y_i \leq y\}\end{aligned}$$

where $N_{\mathbf{x}}$ is the number of observations i where

$$x_{i1}=x_1, x_{i2}=x_2, \cdots, x_{ik}=x_k.$$

For compactness, we'll also let $N_{\mathbf{x}}$ represent the set of all i where the above condition is true.

An Analogy Estimator for the CEF

Using the empirical conditional distribution function, we can show as before that

$$\begin{aligned}
\widehat{\mu}_{y|\mathbf{x}} &= \int_y y d\widehat{D}(y|\mathbf{x}) \\
&= \sum_{i \in N_{\mathbf{x}}} y_i \left[\widehat{D}^+(y_i|\mathbf{x}) - \widehat{D}^-(y_i|\mathbf{x}) \right] \\
&= \frac{1}{N_{\mathbf{x}}} \sum_{i \in N_{\mathbf{x}}} y_i \\
&= \overline{y}_{N_{\mathbf{x}}}
\end{aligned}$$

so, when $\mathbf{x}$ contains only discrete values, the sample means of y_i conditional on $i \in N_{\mathbf{x}}$ are our analogy estimator of the CEF.

An Analogy Estimator for the CEF

For example, if

- x_{i1} represents i's sex (coded 0/1)
- x_{i2} represents i's race (coded 0/1/2),

then $\mathbf{x} = [0, 1]$ represents conditioning on the observed covariates $x_1 = 0$ (male) and $x_2 = 1$ (black).

$N_{0,1}$ counts (and represents) individuals for whom $x_{i1} = 0$ and $x_{i2} = 1$ (i.e., black males) in the sample.

In this case, our estimate of the CEF consists of six sample means, corresponding to the six sets of observations: $N_{0,0}, N_{0,1}, \ldots, N_{1,2}$.

An Analogy Estimator for the CEF

As the number of discrete covariates grows, $\mathbf{x}$ gets longer and the number of crosstabs implied by $\mathbf{x}$ grows.

As the number of crosstabs grows, $N_{\mathbf{x}}$ falls, so our cell counts will shrink. Our estimates of the sample mean from these smaller subsamples (conditional sample means) will become less and less precise.

The situation becomes impossible when $\mathbf{x}$ contains continuous variables.

CPS Examples using OLS and Conditional Sample Means

Now we will work through some examples, examining the joint distribution of log wages and several sets of observed covariates.

CPS: Sample Mean of Log Wage by Sex

This is the analog estimator for the CEF ($\mu_{y|\mathbf{x}}$).

```
> # Recode sex as "factor" variable
> cps$sex <- factor(cps$female,labels=c("Male","Female"))
```

Using built-in tools:

```
> # Show information by sex
> with(cps,tapply(lwage,sex,mean))
    Male   Female
3.045938 2.811624
> with(cps,tapply(lwage,sex,function(v) length(v)))
  Male Female
 29140  21602
```

Using `dplyr`:

```
> suppressMessages(library(dplyr))
> cps %>% group_by(sex) %>% summarize(mean(lwage),n())
# A tibble: 2 x 3
  sex    `mean(lwage)` `n()`
  <fct>          <dbl> <int>
1 Male            3.05 29140
2 Female          2.81 21602
```

CPS: Empirical Conditional DF of Log Wage by Sex

Plot the empirical conditional distribution function $\widehat{D}(\text{lwage}|\text{sex})$

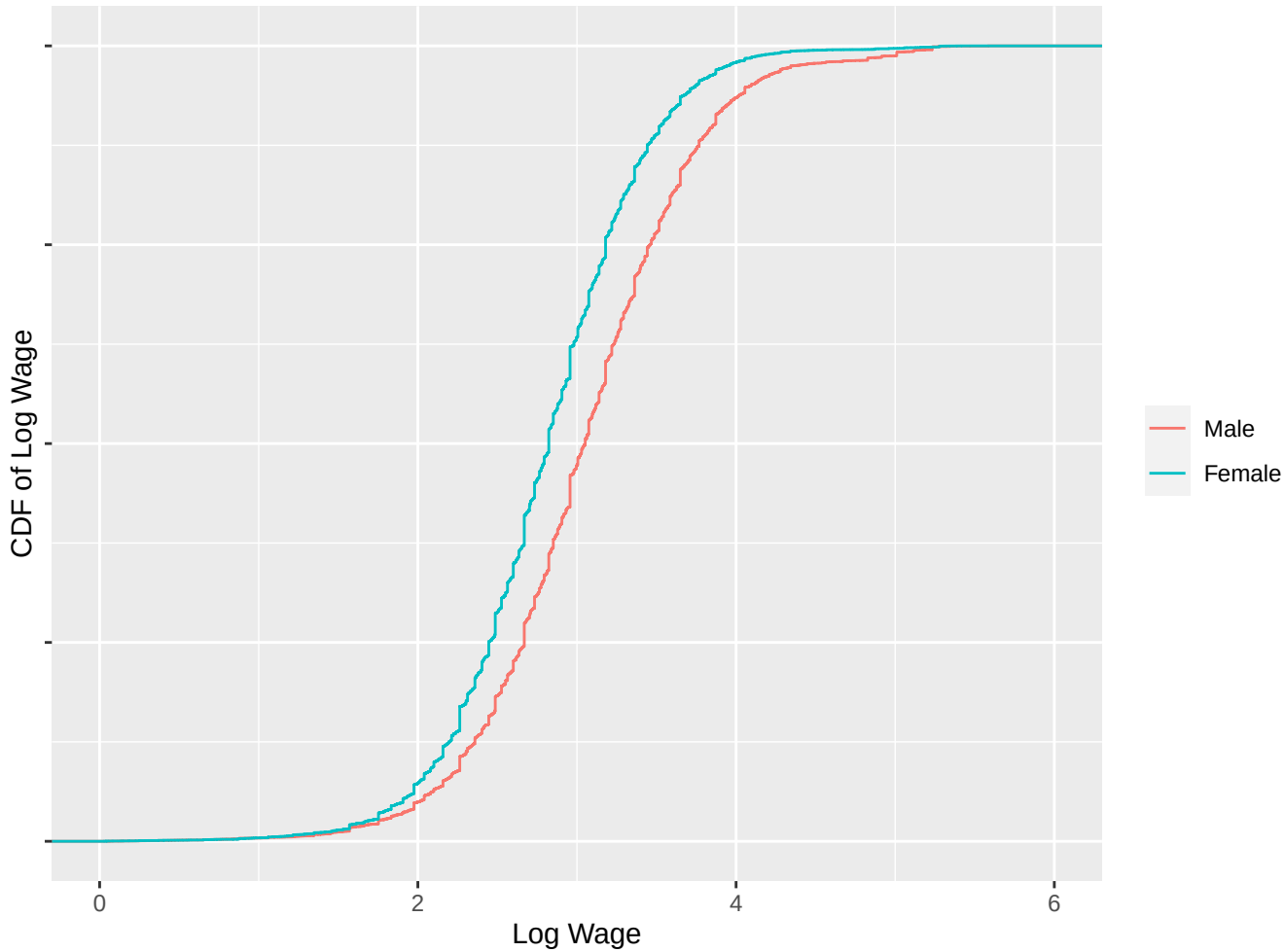
- sex $= 0$ ("Male", red), and
- sex=1 ("Female",blue)

```
> # More information than the mean: plot the empirical d.f. of log wage by sex
> library(ggplot2)
> cdf.lwage.by.sex <- ggplot(cps,aes(x=lwage,color=sex)) +
+     stat_ecdf(geom="step",na.rm=T) +
+     xlim(0,6) +
+     xlab("Log Wage") +
+     ylab("CDF of Log Wage") +
+     theme(
+     axis.text.y=element_blank(),
+         legend.title=element_blank()
+     ) +
+     coord_fixed(ratio=5)
```

I have switched from R's "base graphics" to "grid graphics" (`ggplot2`) for these plots.

CPS: Empirical Conditional DF of Log Wage by Sex

```
> print(cdf.lwage.by.sex)
```



CPS: Empirical Density of Log Wage by Sex

```
> # More information than the mean: plot empirical density of log wage by sex
> pdf.lwage.by.sex <- ggplot(cps,aes(x=lwage,fill=sex)) +
+     geom_density(color=NA,alpha=.4,bw=.08,na.rm=T) +
+     xlim(0,6) +
+     xlab("Log Wage") +
+     ylab("Density of Log Wage") +
+     theme(
+         axis.text.y=element_blank(),
+         legend.title=element_blank()
+     ) +
+     coord_fixed(ratio=5)
```

CPS: Empirical Density of Log Wage by Sex

```
> print(pdf.lwage.by.sex)
```

CPS: Mean Log Wage by Sex in Regression Form

This is the analog estimator for the LP ($\mathbb{L}[y|\mathbf{x}]$)

```
> res.0 <- lm(lwage~0+sex,data=cps); summary(res.0)

Call:
lm(formula = lwage ~ 0 + sex, data = cps)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9092  -0.3771   0.0100   0.3940   2.6019

Coefficients:
          Estimate Std. Error t value Pr(>|t|)
sexMale   3.045938   0.003901   780.8   <2e-16 ***
sexFemale 2.811624   0.004531   620.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6659 on 50740 degrees of freedom
Multiple R-squared:  0.9515,Adjusted R-squared:  0.9515
F-statistic: 4.974e+05 on 2 and 50740 DF,  p-value: < 2.2e-16
```

Why are the results from the LP and the CEF estimators equal?

CPS: Log Wage by Sex: CEF vs. LP

Both the CEF ($\mathbb{E}[y|\mathbf{x}]$) and the LP ($\mathbb{L}[y|\mathbf{x}]$) are functions of $\mathbf{x}$, so one natural question is how different values of $\mathbf{x}$ relate to different best-predictions $\mathbb{E}[y|\mathbf{x}]$ or $\mathbb{L}[y|\mathbf{x}]$.

In the case where $\mathbf{x}$ consists of a single indicator ("dummy") variable or a set of indicator variables built from a single categorical variable, the relationship is best characterized by differencing.

So for `sex`, which is in this data set a binary indicator variable, we have constructed $\mathbf{x}_0$ and $\mathbf{x}_1$ as mutually orthogonal and exhaustive indicators, i.e., $\mathbf{x}_0'\mathbf{x}_1 = 0$ and $\mathbf{x}_0 + \mathbf{x}_1 = \iota$.

CPS: Log Wage by Sex: Marginal Effects

For the CEF, the difference in mean log wage between men and women is

$$\widehat{\mu}_{y|x_0=1} - \widehat{\mu}_{y|x_1=1} = 3.046 - 2.812 = 0.234$$

For the LP, the corresponding value is

$$\mathbb{L}[y|\text{sex} = \text{Male}] - \mathbb{L}[y|\text{sex} = \text{Female}]$$

We can get R to predict out these values from our model automatically using `predict`:

```
> predict(res.0,data.frame(sex="Male")); predict(res.0,data.frame(sex="Female"))
       1
3.045938
       1
2.811624
> predict(res.0,data.frame(sex="Male"))-predict(res.0,data.frame(sex="Female"))
        1
0.2343138
```

Estimates from the LP are identical to those from the CEF.

CPS: Log Wage by Sex: Marginal Effects

We can make things even easier on ourselves by including a constant term in the regression and only one of the indicator variables for `sex`:

```
> res.1 <- lm(lwage~sex,data=cps); summary(res.1)

Call:
lm(formula = lwage ~ sex, data = cps)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9092  -0.3771   0.0100   0.3940   2.6019

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.045938   0.003901  780.84   <2e-16 ***
sexFemale   -0.234314   0.005979  -39.19   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6659 on 50740 degrees of freedom
Multiple R-squared:  0.02938,Adjusted R-squared:  0.02936
F-statistic:  1536 on 1 and 50740 DF,  p-value: < 2.2e-16
```

Now the coefficient on `sex` is $-\delta$ in the LP.

CPS: Log Wage by Sex and Race

```
> # Recode race as factor to match Hansen p. 16
> cps$race3 <- cut(cps$race,breaks=c(0,1,2,Inf),labels=c("White","Black","Other"))
```

Using built-in tools:

```
> # Show information by sex
> with(cps,tapply(lwage,list(sex,race3),mean))
          White    Black    Other
Male   3.065787 2.864154 3.027230
Female 2.819883 2.726742 2.858445
> with(cps,tapply(lwage,list(sex,race3),function(v) length(v)))
       White Black Other
Male   24344  2413  2383
Female 16932  2722  1948
```

Using `dplyr`:

```
> cps %>% group_by(sex,race3) %>% summarize(mean(lwage),n())
`summarise()` has grouped output by 'sex'.  You can override using the `.groups`
argument.
# A tibble: 6 x 4
# Groups:   sex [2]
  sex    race3 `mean(lwage)` `n()`
  <fct>  <fct>         <dbl> <int>
1 Male   White          3.07 24344
2 Male   Black          2.86  2413
3 Male   Other          3.03  2383
4 Female White          2.82 16932
5 Female Black          2.73  2722
6 Female Other          2.86  1948
```

Log Wage by Sex and Race in Regression Form

```
> res.2p <- lm(lwage~sex+race3,data=cps); summary(res.2p)

Call:
lm(formula = lwage ~ sex + race3, data = cps)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9217  -0.3850  -0.0029   0.3884   2.6710

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.058425   0.004089 747.999   <2e-16 ***
sexFemale   -0.227956   0.005982 -38.106   <2e-16 ***
race3Black  -0.146274   0.009858 -14.837   <2e-16 ***
race3Other  -0.004581   0.010616  -0.431    0.666
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6645 on 50738 degrees of freedom
Multiple R-squared:  0.0336,Adjusted R-squared:  0.03354
F-statistic:   588 on 3 and 50738 DF,  p-value: < 2.2e-16
```

What explains why the coefficient on (White,Male) is 3.058 here and 3.065 on the previous slide? That is, why are the results from the LP and the CEF estimators **NOT** equal?

Log Wage by Sex and Race

```
> res.2f <- lm(lwage~sex*race3,data=cps); summary(res.2f)

Call:
lm(formula = lwage ~ sex * race3, data = cps)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9291  -0.3812   0.0031   0.3925   2.7190

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.065787   0.004257 720.164  < 2e-16 ***
sexFemale           -0.245905   0.006647 -36.997  < 2e-16 ***
race3Black          -0.201633   0.014176 -14.224  < 2e-16 ***
race3Other          -0.038557   0.014257  -2.704 0.006844 **
sexFemale:race3Black  0.108493   0.019725   5.500 3.81e-08 ***
sexFemale:race3Other  0.077119   0.021349   3.612 0.000304 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6642 on 50736 degrees of freedom
Multiple R-squared:  0.03435,Adjusted R-squared:  0.03426
F-statistic:   361 on 5 and 50736 DF,  p-value: < 2.2e-16
```

What does the formula operator "$*$" do? Why are the standard errors on the last four coefficients so much higher than the first two?

CPS: Log Wage by Sex and Race: Marginal Effects

As with the previous example, the CEF and LP for the full model are identical, so focus on the difference between the CEF and the LP for the partial model (LP^p). Define

```
> expand.grid(sex=c("Male","Female"),race3=c("White","Black"))
     sex race3
1   Male White
2 Female White
3   Male Black
4 Female Black
> CEF.vals <- predict(res.2f,expand.grid(sex=c("Male","Female"),race3=c("White","Black")))
> CEF.vals[1]-CEF.vals[2]
        1
0.2459046
> CEF.vals[3]-CEF.vals[4]
        3
0.1374116
```

So

$$\mathbb{E}[y|\text{sex=M,race=W}] - \mathbb{E}[y|\text{sex=F,race=W}] = 0.2459046$$
$$\mathbb{E}[y|\text{sex=M,race=B}] - \mathbb{E}[y|\text{sex=F,race=B}] = 0.1374116$$

CPS: Log Wage by Sex and Race: Marginal Effects

Similarly, for the LP in the partial model define

```
> LPP.vals <- predict(res.2p,expand.grid(sex=c("Male","Female"),race3=c("White","Black")))
> LPP.vals
       1        2        3        4
3.058425 2.830468 2.912151 2.684194
```

So

$$LP^p[y|\text{sex=M,race=W}] - LP^p[y|\text{sex=F,race=W}] = 0.2279563$$
$$LP^p[y|\text{sex=M,race=B}] - LP^p[y|\text{sex=F,race=B}] = 0.2279563$$

Evidently, these two differences need not be equal as they are in the LP^p. In the CEF they are not. The missing interaction forces the LP^p away from the CEF.

CPS: Log Wage by Sex and Race: Marginal Effects

Another way to say this is that, in the LP for the full model, the difference between average men's and women's wages may depend on race, or

$$\begin{aligned} LP^f[y|\text{sex,race}] = \beta_1 + \beta_2 \cdot \mathbf{1}\{\text{sex=F}\} + \\ \beta_3 \cdot \mathbf{1}\{\text{race=B}\} + \beta_4 \cdot \mathbf{1}\{\text{race=O}\} \\ \beta_5 \cdot \mathbf{1}\{\text{sex=F}\} \cdot \mathbf{1}\{\text{race=B}\} \\ \beta_6 \cdot \mathbf{1}\{\text{sex=F}\} \cdot \mathbf{1}\{\text{race=O}\} \end{aligned}$$

So the difference in predicted log wage (y) between men and women is

$$\begin{aligned} LP^f[y|\text{sex=M,race}] - LP^f[y|\text{sex=F,race}] = \\ -\beta_2 - \beta_5 \cdot \mathbf{1}\{\text{race=B}\} - \beta_6 \cdot \mathbf{1}\{\text{race=O}\} \end{aligned}$$

CPS: Log Wage by Sex and Race: Marginal Effects

By contrast, for the partial model

$$LP^p[y|\text{sex,race}] = \gamma_1 + \gamma_2 \cdot \mathbf{1}\{\text{sex=F}\} + \gamma_3 \cdot \mathbf{1}\{\text{race=B}\} + \gamma_4 \cdot \mathbf{1}\{\text{race=O}\}$$

So the difference in predicted log wage (y) between men and women (the "marginal effect" of `sex`) is

$$LP^p[y|\text{sex=M,race}] - LP^p[y|\text{sex=F,race}] = -\gamma_2$$

which is constant.

Minimum Empirical Risk Estimation

Estimating the Minimum Risk (Bayes-Optimal) Hypothesis

- Suppose we want to target the Bayes-optimal hypothesis, as

$$\begin{aligned}\phi(F) &= R^* \\ &= \arg\min_{\theta(\mathbf{x})} R(\theta) \\ &= \arg\min_{\theta(\mathbf{x})} \mathbb{E}_F[\ell(\theta(\mathbf{x}), y)]\end{aligned}$$

- Then we replace F with its empirical distribution, getting

$$\arg\min_{\theta(\mathbf{x})} \widehat{\mathbb{E}}_{\widehat{F}}[\ell(\theta(\mathbf{x}), y)] = \arg\min_{\theta(\mathbf{x})} \frac{1}{N} \sum_{i=1}^{N} \ell(\theta(\mathbf{x}_i), y_i)$$

- The solution to this problem, if it exists, is the minimum empirical risk estimator.

Minimum Empirical Risk Estimation with MSE Loss and Linearity

- Suppose $\theta(\mathbf{x})$ is linear, so $\theta(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$.
- Under MSE loss, the empirical risk minimization problem is given by

$$\begin{aligned}
\arg\min_{\theta(\mathbf{x})} \widehat{\mathbb{E}}_{\widehat{F}}[\ell(\theta(\mathbf{x}), y)] &= \arg\min_{\theta(\mathbf{x})} \frac{1}{N} \sum_{i=1}^{N} \ell(\theta(\mathbf{x}_i), y_i) \\
&= \arg\min_{\theta(\mathbf{x})} \frac{1}{N} \sum_{i=1}^{N} (y_i - \theta(\mathbf{x}_i))^2 \\
&= \arg\min_{\boldsymbol{\beta}} \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{x}_i\boldsymbol{\beta})^2
\end{aligned}$$

Minimum Empirical Risk Estimation with MSE Loss and Linearity

- A differential characterization of the minimum is available by setting the partial derivatives $\frac{\partial}{\partial \beta_j}$ equal to zero for $j = 1, \ldots, m$.
- Doing this gives

$$\frac{\partial}{\partial \beta_j}\left[\frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i\boldsymbol{\beta})^2\right] = \frac{1}{N}\sum_{i=1}^{N} 2(y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}})x_{ij} = 0$$

- The factors of $\frac{1}{N}$ and 2 are irrelevant, so we end up with exactly the same conditions we got for our analogy estimator for the best linear predictor.
- Another way of writing this is that $\sum_{i=1}^{N} u_i x_{ij} = 0$, or $\mathbf{u}^\top \mathbf{x}_{(j)} = 0$ for $j = 1, \ldots, m$. The residuals are orthogonal to the covariates.
- None of this is really surprising, but it works.

Minimum Empirical Risk Estimation with KL Loss and Normality

- Suppose that $\eta = \mathbf{x}\boldsymbol{\beta}$ is linear, but that this time we are fitting F with parametric functions F_θ from a normal EDF.
- We use the natural response (identity) so $\theta(\mathbf{x}) = \eta = \mathbf{x}\boldsymbol{\beta}$.
- Our empirical risk minimization problem is now

$$\begin{aligned}
\arg\min_{\theta(\mathbf{x})} \widehat{\mathbb{E}}_{\widehat{F}}[\ell(\theta(\mathbf{x}), y)] &= \arg\min_{\theta(\mathbf{x})} \frac{1}{N}\sum_{i=1}^{N} \ell(\theta(\mathbf{x}_i), y_i) \\
&= \arg\min_{\theta(\mathbf{x})} \frac{1}{N}\sum_{i=1}^{N} [y_i\theta(\mathbf{x}_i) - b(\theta(\mathbf{x}_i))] \\
&= \arg\min_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N} \left[y_i\mathbf{x}_i\boldsymbol{\beta} - \frac{1}{2}(\mathbf{x}_i\boldsymbol{\beta})^2\right]
\end{aligned}$$

recalling that $b(\theta) = \frac{1}{2}\theta^2$ for the normal EDF.

Minimum Empirical Risk Estimation with KL Loss and Normality

- As before, differential characterization of the minimum is available by setting the partial derivatives $\frac{\partial}{\partial \beta_j}$ equal to zero for $j = 1, \ldots, m$.
- Doing this gives

$$\frac{\partial}{\partial \beta_j}\left[\frac{1}{N}\sum_{i=1}^{N}\left[y_i \mathbf{x}_i \boldsymbol{\beta} - \frac{1}{2}(\mathbf{x}_i \boldsymbol{\beta})^2\right]\right] = \frac{1}{N}\sum_{i=1}^{N}(y_i x_{ij} - \mathbf{x}_i \widehat{\boldsymbol{\beta}} x_{ij})$$
$$= \frac{1}{N}\sum_{i=1}^{N}(y_i - \mathbf{x}_i \widehat{\boldsymbol{\beta}}) x_{ij} = 0$$

  for each $j = 1, \ldots, m$.
- Which is exactly what we got under MSE loss.

Minimum Empirical Risk Estimation with KL Loss and Bernoulli

- Finally, suppose that $\eta = \mathbf{x}\boldsymbol{\beta}$ is linear, but that this time we are fitting F with parametric functions F_θ from a Bernoulli EDF.
- We use the natural response (logistic) so again $\theta(\mathbf{x}) = \eta = \mathbf{x}\boldsymbol{\beta}$.
- Our empirical risk minimization problem is

$$\begin{aligned}
\arg\min_{\theta(\mathbf{x})} \widehat{\mathbb{E}}_{\widehat{F}}[\ell(\theta(\mathbf{x}), y)] &= \arg\min_{\theta(\mathbf{x})} \frac{1}{N}\sum_{i=1}^{N} \ell(\theta(\mathbf{x}_i), y_i) \\
&= \arg\min_{\theta(\mathbf{x})} \frac{1}{N}\sum_{i=1}^{N} [y_i\theta(\mathbf{x}_i) - b(\theta(\mathbf{x}_i))] \\
&= \arg\min_{\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N} \left[y_i\mathbf{x}_i\boldsymbol{\beta} - \log(1 + e^{\mathbf{x}_i\boldsymbol{\beta}})\right]
\end{aligned}$$

recalling that $b(\theta) = \log(1 + e^\theta)$ for the Bernoulli EDF.

Minimum Empirical Risk Estimation with KL Loss and Bernoulli

- As before, differential characterization of the minimum is available by setting the partial derivatives $\frac{\partial}{\partial \beta_j}$ equal to zero for $j = 1, \ldots, m$.
- Doing this gives

$$\frac{\partial}{\partial \beta_j}\left[\frac{1}{N}\sum_{i=1}^{N}\left[y_i \mathbf{x}_i \boldsymbol{\beta} - \log(1 + e^{\mathbf{x}_i \boldsymbol{\beta}})\right]\right] = \frac{1}{N}\sum_{i=1}^{N}\left[y_i x_{ij} - \frac{x_{ij} e^{\mathbf{x}_i \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i \boldsymbol{\beta}}}\right]$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \frac{e^{\mathbf{x}_i \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i \boldsymbol{\beta}}}\right) x_{ij} = 0$$

  for each $j = 1, \ldots, m$.
- These are precisely the first order conditions for the log-likelihood function in logit estimation.
- In this form, it's also clear that the logit "residuals" are being set orthogonal to the covariates.

Estimating Generalized Linear Models in R

- Estimating GLMs in R is quite straightforward, using the built-in function `glm()`.
- We will go through estimation of an instance of each of the three prior examples, estimating
  - a linear model under minimum MSE
  - the same linear model under minimum KL divergence with a normal family (identity link)
  - a linear model for binary outcomes under minimum KL divergence using a Bernoulli family (logit link)

Example: A Log-Wage Regression (MSE)

```
> cps$exp <- cps$age-cps$education-6
> res.mse <- lm(lwage~sex+race3+exp+I(exp^2),data=cps); summary(res.mse)

Call:
lm(formula = lwage ~ sex + race3 + exp + I(exp^2), data = cps)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8795  -0.3716   0.0033   0.3882   2.8559

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.656e+00  1.016e-02 261.306   <2e-16 ***
sexFemale   -2.245e-01  5.872e-03 -38.236   <2e-16 ***
race3Black  -1.464e-01  9.676e-03 -15.127   <2e-16 ***
race3Other   1.295e-02  1.043e-02   1.241    0.214
exp          3.886e-02  8.919e-04  43.570   <2e-16 ***
I(exp^2)    -7.364e-04  1.838e-05 -40.062   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6521 on 50736 degrees of freedom
Multiple R-squared:  0.06918,Adjusted R-squared:  0.06909
F-statistic: 754.2 on 5 and 50736 DF,  p-value: < 2.2e-16
```

Example: A Log-Wage Regression (GLM)

```
> res.glm <- glm(lwage~sex+race3+exp+I(exp^2),family=gaussian(link="identity"),data=cps); summary(res.glm)

Call:
glm(formula = lwage ~ sex + race3 + exp + I(exp^2), family = gaussian(link = "identity"),
    data = cps)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-10.8795   -0.3716    0.0033    0.3882    2.8559

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.656e+00  1.016e-02 261.306   <2e-16 ***
sexFemale   -2.245e-01  5.872e-03 -38.236   <2e-16 ***
race3Black  -1.464e-01  9.676e-03 -15.127   <2e-16 ***
race3Other   1.295e-02  1.043e-02   1.241    0.214
exp          3.886e-02  8.919e-04  43.570   <2e-16 ***
I(exp^2)    -7.364e-04  1.838e-05 -40.062   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.425264)

    Null deviance: 23180  on 50741  degrees of freedom
Residual deviance: 21576  on 50736  degrees of freedom
AIC: 100621

Number of Fisher Scoring iterations: 2
```

Example: A Logit Regression

Modeling labor force participation (`inlf`$= 1$ means in labor force)

```
> library(wooldridge); data(mroz);
> res.logit <- glm(inlf~nwifeinc+educ+exper+expersq+age+kidslt6+kidsge6,
+                  family=binomial(link="logit"),data=mroz)
> summary(res.logit)

Call:
glm(formula = inlf ~ nwifeinc + educ + exper + expersq + age +
    kidslt6 + kidsge6, family = binomial(link = "logit"), data = mroz)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1770  -0.9063   0.4473   0.8561   2.4032

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.425452   0.860365   0.495  0.62095
nwifeinc    -0.021345   0.008421  -2.535  0.01126 *
educ         0.221170   0.043439   5.091 3.55e-07 ***
exper        0.205870   0.032057   6.422 1.34e-10 ***
expersq     -0.003154   0.001016  -3.104  0.00191 **
age         -0.088024   0.014573  -6.040 1.54e-09 ***
kidslt6     -1.443354   0.203583  -7.090 1.34e-12 ***
kidsge6      0.060112   0.074789   0.804  0.42154
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  803.53  on 745  degrees of freedom
AIC: 819.53

Number of Fisher Scoring iterations: 4
```