These notes come from Chapter 3 of the textbook and covers properties of least squares regression under the assumptions of the linear projection model.

Linear Regression Notes 2: The algebra of least squares

We'll spend the next few classes talking about the linear projection model. Recall that, in chapter 2, we defined the **best linear predictor** of Y given X as $X'\beta$ where

$$\beta = \underset{b}{\text{argmin}}\ \text{E}[(Y - X'b)^2]$$

which has the solution $\beta = \text{E}[XX']^{-1}\text{E}[XY]$ (see Hansen 2.18 and our preliminary notes for this derivation though we will derive a very closely related result below).

Moreover, the best linear predictor is therefore given by $X'\beta = X'\text{E}[XX']^{-1}\text{E}[XY]$. And, recall that we defined the **projection error** $e = Y - X'\beta$.

Notation:

H: 3.10

It's convenient to define the **data matrix**

$$\mathbf{X} := \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1k} \\ X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix} = \begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix}$$

which is an $n \times k$ matrix. This is called a data matrix because it is very similar to, say, an Excel spreadsheet — each row contains a particular observation. In class, since it is hard to write bold font, I'll typically use $\underline{X}$ for this data matrix. And, slightly abusing notation, let's define

$$\mathbf{Y} := \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \qquad \text{and} \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

which are both $n \times 1$ vectors. In class, I will write these as $\underline{Y}$ and $\underline{e}$.

Sampling

H: 3.2, 4.2

For most of this semester, we'll suppose that we have access to a **simple random simple**. That is, the observed data consists of $\{(Y_i, X_i)\}_{i=1}^n$ that are independent and identically distributed. This

means that we observe n "draws'' from some underlying population. That the data is identically distributed means that the draws are all from the same distribution — you should not think of this as being a strong assumption in most cases (though there are some exceptions such as some types of time series data). Independent means that observations are independent of each other; for example, if you are studying labor market outcomes, it means that if you draw a very rich person for the first observation, it does not give you an indication of whether or not the next draw is likely to be a rich person or not. Data being independent is a leading case though there are some important exceptions (many time series do not satisfy independence, and we'll talk some about clustered sampling too).

Estimating β from the linear projection model

H: 3.3, 3.4, 3.6

Often in econometrics, we will be interesting in estimating population quantities using the data that we have access. Perhaps the simplest example of a population quantity that we might be interested in estimating is $\mathrm{E}[Y]$. The natural way to estimate a population moment like this is to use its sample analogue. That is,

$$\hat{\mathrm{E}}[Y] := \frac{1}{n}\sum_{i=1}^{n} Y_i$$

where the "$\hat{}$'' indicates that it is an estimated quantity. This strategy of estimating population moments by their sample counterpart is called a **moment estimator** or, more generally, as the **analogy principle**. This strategy also works for estimating more complicated population quantities like the $k \times 1$ matrix $\mathrm{E}[XY]$ or the $k \times k$ matrix $\mathrm{E}[XX']$, where the natural estimators are given by

$$\frac{1}{n}\sum_{i=1}^{n} X_i Y_i \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$$

Another common setup is that we are interested in population parameters that are a function of moments. For example, we can write the linear proejection coefficient $\beta = g(\mathrm{E}[XX'], \mathrm{E}[XY]) = \mathrm{E}[XX']^{-1}\mathrm{E}[XY]$. A natural way to estimate parameters that are functions of moments (as is β) is to use the same function but replace the population moments with their sample counterparts. That is,

$$\begin{aligned}\hat{\beta} &= g\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i', \frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i Y_i\end{aligned}$$

This strategy is called a **plug-in estimators**. This is the way that we'll estimate β, but let's

also think about one other motivation for this expression. We can also think of $\hat{\beta}$ as the solution to

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\beta)^2$$

Let's solve this for $\hat{\beta}$. It's helpful to expand the expression above into

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - 2\beta'\frac{1}{n}\sum_{i=1}^{n} X_iY_i + \beta'\frac{1}{n}\sum_{i=1}^{n} X_iX_i'\beta$$

Now, let's take the derivative with respect to β, set it equal to 0, and solve:

$$0 = -2\frac{1}{n}\sum_{i=1}^{n} X_iY_i + 2\frac{1}{n}\sum_{i=1}^{n} X_iX_i'\hat{\beta}$$

which (as long as $\frac{1}{n}\sum_{i=1}^{n} X_iX_i'$ is positive definite) implies that

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_iX_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^{n} X_iY_i$$

which is exactly the same expression that we got before.
It is also good to have an expression for $\hat{\beta}$ in terms of the data matrices that we defined earlier. Using this notation, notice that we can write

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

which holds because

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix}\begin{pmatrix} X_1' \\ X_2' \\ \vdots \\ X_n' \end{pmatrix} = \sum_{i=1}^{n} X_iX_i'$$

and, similarly,

$$\boldsymbol{X}'\boldsymbol{Y} = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix}\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \sum_{i=1}^{n} X_iY_i$$

Least squares residuals

H: 3.8

Once we have estimated the linear projection coefficient, we can define **fitted values**: $\hat{Y}_i := X_i'\hat{\beta}$

and **residuals**: $\hat{e}_i := Y_i - \hat{Y}_i = Y_i - X_i'\hat{\beta}$. Note that the residual is distinct from the error term that we defined earlier; in particular, the residual is something that we can calculate while the error term is not observed. Two useful properties of residuals are

$$\frac{1}{n}\sum_{i=1}^{n} X_i\hat{e}_i = 0 \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i = 0$$

where the last one holds if X includes an intercept. Proving these results is mechanically similar to proving that $\mathrm{E}[Xe] = 0$ in the linear projection model, so I will omit showing this (you can also see the textbook for a proof).

We can also define a vector/matrix version of the residuals. In particular, define

$$\hat{\boldsymbol{e}} := \boldsymbol{Y} - \boldsymbol{X}\hat{\beta}$$

which is an $n \times 1$ vector of residuals.

Projection Matrix

H: 3.11

Next, let's define

$$\mathbf{P} := \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

$\mathbf{P}$ is called a **projection matrix**. It is an $n \times n$ matrix. Notice that

$$\mathbf{PX} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X}$$

The projection matrix creates fitted values; that is

$$\mathbf{PY} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{X}\hat{\beta} = \hat{\mathbf{Y}}$$

Theorem 3.3 The projection matrix has the following properties (the theorem contains a few extra properties, but these are the main ones we'll use this semester)

1. $\mathbf{P}$ is symmetric
2. $\mathbf{P}$ is idempotent (that is, $\mathbf{PP} = \mathbf{P}$)
3. $\mathrm{tr}(\mathbf{P}) = k$.

Proof

For 1, notice that

$$\begin{aligned}
\mathbf{P}' &= \Big(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Big)' \\
&= (\mathbf{X}')'\Big((\mathbf{X}'\mathbf{X})^{-1}\Big)'(\mathbf{X})' \\
&= \mathbf{X}\left((\mathbf{X}'\mathbf{X})'\right)^{-1}\mathbf{X}' \\
&= \mathbf{X}\left((\mathbf{X})'(\mathbf{X}')'\right)^{-1}\mathbf{X}' \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{P}
\end{aligned}$$

For 2,

$$\begin{aligned}
\mathbf{PP} &= \mathbf{PX}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{P}
\end{aligned}$$

For 3, recall that the trace of a square matrix is the sum of its diagonal elements. It also has the useful property that, for a $k \times r$ matrix $\mathbf{A}$ and an $r \times k$ matrix $\mathbf{B}$, $\text{tr}(AB) = \text{tr}(BA)$ (the proof of this is straightforward, see p.961 in the book).

$$\begin{aligned}
\text{tr}(\mathbf{P}) &= \text{tr}\Big(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Big) \\
&= \text{tr}\Big(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\Big) \\
&= \text{tr}\Big(\mathbf{I}_k\Big) \\
&= k
\end{aligned}$$

A final useful property of $\mathbf{P}$ is that it is positive semi-definite. This holds because all its eigenvalues are either 1 or 0 (and, therefore, non-negative), which implies that it is positive semi-definite.

Annihilator Matrix

H: 3.12

Next, consider the **annihilator matrix**

$$\begin{aligned}
\mathbf{M} &:= \boldsymbol{I}_n - \boldsymbol{P} \\
&= \boldsymbol{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'
\end{aligned}$$

This is also an $n \times n$ matrix. Notice that,

$$\begin{aligned}\mathbf{MX} &= (\boldsymbol{I}_n - \mathbf{P})\mathbf{X} \\ &= \mathbf{X} - \mathbf{X} \\ &= \mathbf{0}\end{aligned}$$

where $\mathbf{0}$ is an $n \times k$ matrix of zeros.

Further, $\mathbf{M}$ creates residuals (in fact, it is sometimes called the residual-maker matrix). To see this, notice that

$$\begin{aligned}\mathbf{MY} &= (\boldsymbol{I}_n - \mathbf{P})Y \\ &= \mathbf{Y} - \mathbf{PY} \\ &= \mathbf{Y} - \mathbf{X}\hat{\beta} \\ &= \hat{\mathbf{e}}\end{aligned}$$

We can also use the annihilator matrix to provide a useful expression for the residuals in terms of the linear projection errors. In particular, notice that

$$\begin{aligned}\hat{\mathbf{e}} &= \mathbf{MY} \\ &= \mathbf{M}(\mathbf{X}\beta + \mathbf{e}) \\ &= \mathbf{0} + \mathbf{Me} \\ &= \mathbf{Me}\end{aligned}$$

The annihilator matrix has some similar properties as the projection matrix above. In particular, it is symmetric, idempotent, and $\text{tr}(\mathbf{M}) = n - k$. I'll omit the proofs of these, but it is good practice to show that these statements are true.

Estimation of Error Variance

H: 3:13

Next, let's consider trying to estimate the variance of the linear projection error; that is $\sigma^2 := \text{E}[e^2]$. Following our earlier discussion, if the error term were observed for particular observations, we would estimate it by

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$$

This is an infeasible estimator though (because e_i is not observed). A main alternative way to

estimate σ^2 is the feasible estimator that replaces the error term, e_i, with the residual $\hat{e}_i$. That is,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{e}_i^2$$

Let's try to relate this to the infeasible estimator above. Notice that,

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\hat{\mathbf{e}}'\hat{\mathbf{e}} \\
&= \frac{1}{n}(\mathbf{Me})'\mathbf{Me} \\
&= \frac{1}{n}\mathbf{e}'\mathbf{M}'\mathbf{Me} \\
&= \frac{1}{n}\mathbf{e}'\mathbf{MMe} \\
&= \frac{1}{n}\mathbf{e}'\mathbf{Me}
\end{aligned}$$

where the first equality holds by the definition of $\hat{\mathbf{e}}$, the second equality holds by our earlier argument that $\hat{\mathbf{e}} = \mathbf{Me}$, the third equality holds by taking the transpose, the fourth equality holds because $\mathbf{M}$ is symmetric, the fifth equality holds because $\mathbf{M}$ is idempotent. This implies that

$$\begin{aligned}
\tilde{\sigma}^2 - \hat{\sigma}^2 &= \frac{1}{n}\mathbf{e}'\mathbf{e} - \frac{1}{n}\mathbf{e}'\mathbf{Me} \\
&= \frac{1}{n}\mathbf{e}'\boldsymbol{I}_n\mathbf{e} - \frac{1}{n}\mathbf{e}'\mathbf{Me} \\
&= \frac{1}{n}\mathbf{e}'\mathbf{Pe} \geq 0
\end{aligned}$$

where the last equality holds because $\mathbf{P}$ is positive semi-definite and $\mathbf{e}'\mathbf{Pe}$ is a quadratic form. This implies that $\hat{\sigma}^2$ is smaller than the infeasible $\tilde{\sigma}^2$.

Regression Components

H: 3.16

There are a large number of cases where we may be more interested in some of the regression parameters than others (e.g., the treatment effects discussion that we had earlier this semester), so it's useful to have some specific expressions for subsets of the parameters. For this, let's partition $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix}$ and, likewise, $\beta = (\beta_1', \beta_2')'$. Using this notation, we can immediately write

$$\begin{aligned}
\mathbf{Y} &= \mathbf{X}\hat{\beta} + \hat{\mathbf{e}} \\
&= \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \hat{\mathbf{e}}
\end{aligned}$$

Recall that, $\hat{\beta}_1$ and $\hat{\beta}_2$ minimize the sum of squared residuals (we'll write this in a slightly different way for now)

$$(\hat{\beta}_1', \hat{\beta}_2')' = \underset{\beta_1, \beta_2}{\text{argmin}} \ (\mathbf{Y} - \mathbf{X}\beta_1 - \mathbf{X}\beta_2)'(\mathbf{Y} - \mathbf{X}\beta_1 - \mathbf{X}\beta_2)$$

If we are just focused on $\hat{\beta}_1$, we can alternatively express this as

$$\hat{\beta}_1 = \underset{\beta_1}{\text{argmin}} \ \Big(\min_{\beta_2}(\mathbf{Y} - \mathbf{X}\beta_1 - \mathbf{X}\beta_2)'(\mathbf{Y} - \mathbf{X}\beta_1 - \mathbf{X}\beta_2)\Big) \tag{1}$$

This sort of nested minimization is often referred to as "concentrating out'' β_2 and is a fairly common estimation strategy (it doesn't really apply here, but there are some cases where this sort of step may lead to estimators that are notably less computationally complex).

Let's focus on the inside minimization first. For the inside minimization, we treat β_1 as being fixed and the value of β_2 that minimizes this expression will be a function of β_1. This amounts to just a regression of $\mathbf{Y} - \mathbf{X}_1\beta_1$ on $\mathbf{X}_2$ which implies that $\beta_2(\beta_1) = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X_2}'(\mathbf{Y} - \mathbf{X}_1\beta_1)$. The residuals from this regression are given by

$$\begin{aligned}
&\mathbf{Y} - \mathbf{X}_1\beta_1 - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X_2}'(\mathbf{Y} - \mathbf{X}_1\beta_1) \\
&\quad = (\boldsymbol{I}_n - \mathbf{P}_2)(\mathbf{Y} - \mathbf{X}_1\beta_1) \\
&\quad = \mathbf{M}_2(\mathbf{Y} - \mathbf{X}_1\beta_1)
\end{aligned}$$

where $\mathbf{P}_2 := \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X_2}'$ and $\mathbf{M}_2 := (\boldsymbol{I}_n - \mathbf{P}_2)$. Therefore, the inside term in Equation 1 can be written as

$$\begin{aligned}
\min_{\beta_2}(\mathbf{Y} - \mathbf{X}\beta_1 - \mathbf{X}\beta_2)'(\mathbf{Y} - \mathbf{X}\beta_1 - \mathbf{X}\beta_2) &= (\mathbf{M}_2(\mathbf{Y} - \mathbf{X}_1\beta_1))'(\mathbf{M}_2(\mathbf{Y} - \mathbf{X}_1\beta_1)) \\
&= (\mathbf{Y} - \mathbf{X}_1\beta_1)'\mathbf{M}_2(\mathbf{Y} - \mathbf{X}_1\beta_1)
\end{aligned}$$

where the first equality uses the expression for the residuals above and the last equality uses that $\mathbf{M}_2$ is symmetric and idempotent. Now, let's plug this into the outside minimization problem.

$$\begin{aligned}
\hat{\beta}_1 &= \underset{\beta_1}{\text{argmin}} \ (\mathbf{Y} - \mathbf{X}_1\beta_1)'\mathbf{M}_2(\mathbf{Y} - \mathbf{X}_1\beta_1) \\
&= \underset{\beta_1}{\text{argmin}} \ \mathbf{Y}'\mathbf{M}_2\mathbf{Y} - 2\beta_1'\mathbf{X}_1'\mathbf{M}_2\mathbf{Y} + \beta_1'\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\beta_1
\end{aligned}$$

Taking the derivative of the right hand side and setting equal to 0, we have that

$$0 = -2\mathbf{X}_1'\mathbf{M}_2\mathbf{Y} + 2\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\hat{\beta}_1$$

which implies that

$$\hat{\beta}_1 = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y}$$

The arguments above are symmetric, so you could make the same sorts of calculations and derive a similar result for $\hat{\beta}_2$.

Residual Regression

H: 3.18

The previous result is very closely related to a famous result in econometrics called the Frisch, Waugh, Lovell Theorem. In particular, from the previous expression for $\hat{\beta}_1$, we have that

$$\begin{aligned}
\hat{\beta}_1 &= (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{Y} \\
&= (\mathbf{X}_1'\mathbf{M}_2'\mathbf{M}_2\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{M_2}'\mathbf{M}_2\mathbf{Y} \\
&= \left((\mathbf{M}_2\mathbf{X}_1)'\mathbf{M}_2\mathbf{X}_1\right)^{-1}(\mathbf{M}_2\mathbf{X}_1)'\mathbf{M}_2\mathbf{Y} \\
&= (\tilde{\mathbf{X}}_1'\tilde{\mathbf{X}}_1)^{-1}\tilde{\mathbf{X}}_1\tilde{\mathbf{e}}_2
\end{aligned}$$

which uses that $\mathbf{M}_2$ is symmetric and idempotent and where $\tilde{\mathbf{X}}_1 := \mathbf{M}_2\mathbf{X}_1$ (i.e., the residuals from a regression of $\mathbf{X}_1$ on $\mathbf{X}_2$) and $\tilde{\mathbf{e}}_2 := \mathbf{M}_2\mathbf{Y}$ (i.e., the residuals from the regression of $\mathbf{Y}$ on $\mathbf{X}_2$).

This implies an algebraic equivalence between $\hat{\beta}_1$ from the regression of $\mathbf{Y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$ and the following estimation procedure:

1. Regress $\mathbf{Y}$ on $\mathbf{X}_2$ and recover the residuals $\tilde{\mathbf{e}}_2$.
2. Regress $\mathbf{X}_1$ on $\mathbf{X}_2$ and recover the residuals $\tilde{\mathbf{X}}_1$.
3. Regress $\tilde{\mathbf{e}}_2$ on $\tilde{\mathbf{X}}_1$.

This procedure delivers exactly the same estimate of $\hat{\beta}_1$. That this procedure recovers exactly the same estimate of $\hat{\beta}_1$ is called the Frisch-Waugh-Lovell Theorem.

This result gives a nice interpretation to the estimates of $\hat{\beta}_1$. It is equivalent to a regression of $\mathbf{Y}$ on $\mathbf{X}_1$ after "partialling out'' (i.e., removing the effect of $\mathbf{X}_2$ on both $\mathbf{Y}$ and $\mathbf{X}_1$). Besides that, the FWL Theorem is computationally useful in some important cases too such as some of the panel data approaches that we'll consider later in the semester.

Side-Comment:

H: 2.23

A population version of FWL is given in H: 2.23. I have used this in a few of my papers, so I just want to quickly mention it here. For simplicity (and because it is the leading case), let's consider the case where X_1 is scalar and write

$$Y = X_1\beta_1 + X_2'\beta_2 + e$$

where $\mathrm{E}[Xe] = 0$. Now, consider the projection of X_1 on X_2, that is,

$$X_1 = X_2'\gamma_1 + u_1$$

where $\mathrm{E}[X_2 u_1] = 0$. Now, notice that

$$\begin{aligned}
\mathrm{E}[u_1 Y] &= \mathrm{E}[u_1 X_1]\beta_1 + \underbrace{\mathrm{E}[u_1 X_2']}_{=0}\beta_2 + \mathrm{E}[u_1 e] \\
&= \mathrm{E}[u_1(X_2'\gamma_1 + u_1)]\beta_1 + \mathrm{E}[(X_1 - X_2'\gamma_1)e] \\
&= \mathrm{E}[u_1^2]\beta_1
\end{aligned}$$

where the second equality holds by substituting for X_1 in the first term and for u_1 in the last term, and the last equality holds because $\mathrm{E}[X_2 u_1] = 0$ and because $\mathrm{E}[X_1 e] = 0$ and $\mathrm{E}[X_2 e] = 0$. This implies that

$$\beta_1 = \frac{\mathrm{E}[u_1 Y]}{\mathrm{E}[u_1^2]}$$

which is, essentially, a population version of the FWL theorem.

As a final comment, we can write the linear projection of Y on X_2

$$Y = X_2'\gamma_Y + u_Y$$

Using similar arguments as above, you can also show that

$$\beta_1 = \frac{\mathrm{E}[u_Y X_1]}{\mathrm{E}[u_1^2]} \qquad \text{and} \qquad \beta_1 = \frac{\mathrm{E}[u_1 u_Y]}{\mathrm{E}[u_1^2]}$$

but I will leave these as practice problems for you (note that the last expression is really the most analogous version of a population version of FWL).