Linear Regression Mechanics Notes

Exam Technique

- In giving descriptive interpretations of regression parameters, avoid the words "increase" and "decrease" which suggests a change from one state to another rather than a mere comparison between two instances. The general template for such interpretations is as follows. "On average, an [object's] having one unit higher [independent variable] is associated with having [coefficient magnitude] [higher/lower] [dependent variable], holding each other regressor constant."
    - For example, write "women who consumed one more cigarette a day during pregnancy, on average, had infants with $15g$ lower birthweight, holding each other regressor constant."
- The general template for a discussion of whether some regression coefficient can be given a causal interpretation is as follows. "The sample linear regression coefficient is consistent for the causal effect of interest only if [independent variable] is uncorrelated with other unobserved determinants of [dependent variable] such that the abbreviated causal model coincides with the population regression model. The sample regression coefficient is consistent for the coefficient in the population regression model. [Independent variable] is not plausibly uncorrelated with the unobserved determinants of [dependent variable]. One such determinant is [...]. This is likely to be correlated with [independent variable] because [...]."
    - Additionally, the estimated coefficient provides a consistent estimate of the causal effect of interest only if the causal effect of interest is linear.

Simple Linear Regression

Causal Model

- A causal model is a model of some outcome y as a function $y(x_1, x_2, \ldots)$ of a number of determinants $x_1, x_2, \ldots$. In what follows, it is supposed for simplicity that the causal model is linear, i.e. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$.
- For simple linear regression (linear regression of one dependent variable on one independent variable), the causal model simplifies to $y = \beta_0 + \beta_1 x_1 + u$, where $u := \beta_2 x_2 + \ldots$ collects the remaining terms in the causal model of y.
    - Note that $y, x_1, x_2, \ldots$ are used to denote the variables of the causal model because the causal model is a deterministic (as opposed to stochastic) model that accepts non-random inputs, and in such cases yields non-random outputs.
- The causal model, evaluated at some arbitrary observation $(Y, X_1, X_2, \ldots)$ is $Y = \beta_0 + \beta_1 X + u$. X is written in place of X_1 for notational convenience.
    - The causal model also accepts random inputs, and observations (in the population, or in a sample) are random.

Population Regression

- The population linear regression problem is the problem of constructing the best linear predictor of Y given only X. Formally, this problem is $\min_{b_0,b_1} C(b_0, b_1) := \mathbb{E}(Y - b_0 - b_1 X)^2$, where $C(b_0, b_1)$ (for cost) is the mean squared prediction error. Denote the solution to the population linear regression problem as $\rho_0, \rho_1 := \arg\min_{b_0,b_1} C(b_0, b_1)$.
- Solve by taking first-order conditions.
- $$\begin{aligned} FOC_{b_0} : \frac{\partial}{\partial b_0} C(b_0, b_1) &= \frac{\partial}{\partial b_0} \mathbb{E}(Y - b_0 - b_1 X)^2 \\ &= \mathbb{E}[\frac{\partial}{\partial b_0}(Y - b_0 - b_1 X)^2] \\ &= \mathbb{E}[2(Y - b_0 - b_1 X)(-1)] \\ &= 0 \\ \Rightarrow \mathbb{E}(Y - b_0 - b_1 X) = 0 & \end{aligned}.$$
- $$\begin{aligned} FOC_{b_1} : \frac{\partial}{\partial b_1} C(b_0, b_1) &= \frac{\partial}{\partial b_1} \mathbb{E}(Y - b_0 - b_1 X)^2 \\ &= \mathbb{E}[\frac{\partial}{\partial b_1}(Y - b_0 - b_1 X)^2] \\ &= \mathbb{E}[2(Y - b_0 - b_1 X)(-X)] \\ &= 0 \\ \Rightarrow \mathbb{E}[X(Y - b_0 - b_1 X)] = 0 & \end{aligned}.$$
- $\Rightarrow b_0 = \mathbb{E}Y - b_1 \mathbb{E}X$.

  - $$\begin{aligned}\Rightarrow \mathbb{E}[X(Y-b_0-b_1X)] &= \mathbb{E}[X(Y-(\mathbb{E}Y-b_1\mathbb{E}X)-b_1X)] \\ &= \mathbb{E}[X(Y-\mathbb{E}Y-b_1(X-\mathbb{E}X))] \\ &= \mathbb{E}[X(Y-\mathbb{E}Y)]-b_1\mathbb{E}[X(X-\mathbb{E}X)] \\ \Rightarrow b_1 &= \frac{\mathbb{E}[X(Y-\mathbb{E}Y)]}{\mathbb{E}[X(X-\mathbb{E}X)]} \\ &= \frac{\mathbb{E}[XY-X\mathbb{E}Y]}{\mathbb{E}[X^2-X\mathbb{E}X]} = \frac{\mathbb{E}XY-\mathbb{E}X\mathbb{E}Y}{\mathbb{E}X^2-\mathbb{E}X\mathbb{E}X} \\ &= \frac{cov(X,Y)}{var(X)}\end{aligned}.$$
  - $\Rightarrow \rho_0 = \mathbb{E}Y - \rho_1\mathbb{E}X, \rho_1 = \frac{cov(X,Y)}{var(X)}$.
- Let the error term $e := Y - \rho_0 - \rho_1 X$. Then, $\mathbb{E}e = \mathbb{E}Xe = 0$.
- $Y = \rho_0 + \rho_1 X + e$ is the unique decomposition of Y into a linear function of X with an error term e that has mean zero and is orthogonal to X (i.e, $\mathbb{E}Xe = 0$).
- The population linear regression model is the function $Y = \rho_0 + \rho_1 X + e$, and its parameters are ρ_0, ρ_1, which are equivalently characterised by each of the following:
  - $\rho_0, \rho_1 := \arg\min_{b_0,b_1} C(b_0, b_1)$,
  - $\rho_0 = \mathbb{E}Y - \rho_1\mathbb{E}X, \rho_1 = \frac{cov(X,Y)}{var(X)}$, and
  - $Y = \rho_0 + \rho_1 X + e$, where $\mathbb{E}e = \mathbb{E}Xe = 0$.
- The population linear regression model, as such, is not a causal model of Y, but merely a descriptive model that describes Y as an optimal (mean squared error minimising) linear function of X.
- Suppose that orthogonality holds in the causal model $Y = \beta_0 + \beta_1 X + u$, i.e. $\mathbb{E}u = \mathbb{E}Xu = 0$. Note that $Y = \rho_0 + \rho_1 X + e$ is a unique decomposition of Y into a linear function of X that satisfies orthogonality. Then, ρ_0, ρ_1, e coincide with β_0, β_1, u, and $\beta_0 = \mathbb{E}Y - \beta_1\mathbb{E}X, \beta_1 = \frac{cov(X,Y)}{var(X)}$.

Regression and Conditional Expectation

- The conditional expectation $\mathbb{E}[Y|X]$ of Y given X solves $\min_m \mathbb{E}[Y - m(X)]^2$, where m is a potentially non-linear function of X. The conditional expectation coincides with the (population) regression function $y = \rho_0 + \rho_1 x$ iff the conditional expectation is linear.
- Otherwise, the regression function can be interpreted as a linear approximation of the conditional expectation. Then, supposing that the regression function constitutes a reasonable approximation of the conditional expectation, ρ_1 can be interpreted as the derivative of the conditional expectation $\mathbb{E}[Y|X = x]$ with respect to x.
- Mean independence implies that the conditional expectation is linear hence that the regression function coincides with the conditional expectation.
  - $$\begin{aligned}\mathbb{E}[Y|X] &=_1 \mathbb{E}[\beta_0 + \beta_1 X + u|X] \\ &= \beta_0 + \beta_1\mathbb{E}[X|X] + \mathbb{E}[u|X]. \\ &= \beta_0 + \beta_1 X\end{aligned}$$
  - Where $=_1$ follows by substitution of the causal model.

Orthogonality

- Random variable u is orthogonal to random variable X (and u is mean zero) iff $\mathbb{E}u = \mathbb{E}Xu = 0$, which implies $cov(X, u) = 0$.
- Random variable u is mean independent of random variable X (and u is mean zero) iff $\mathbb{E}(u|X) = \mathbb{E}u = 0$.
- Random variable u is independent of random variable X (and u is mean zero) iff $\mathbb{E}u = 0$ and $u \perp\!\!\!\perp X$.
- Independence (of u on X) implies mean independence, which in turn implies orthogonality. Orthogonality is the weakest of the three conditions, and independence is the strongest.
- Orthogonality is sufficient for the estimation of the parameters of the causal model by linear regression. Intuitively, if orthogonality fails, then the causal effect of X cannot be isolated from the statistical effect of X. The statistical effect of X captures both the direct effect of X and the direct effect of other causal factors that are correlated with X.
- If orthogonality fails, the population linear regression model remains well-defined and can be interpreted as a descriptive model that describes Y as an optimal linear function of X; the sample linear regression model also remains well-defined and consistently estimates the population regression parameters.

Sample Regression

- The sample linear regression problem is the problem of constructing the best linear predictor of Y given only X, where (Y, X) here is some arbitrary observation within a given sample $\{(Y_i, X_i)\}_{i=1}^n$.

- The sample linear regression model is the function $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{u}$, and its parameters are $\hat{\beta}_0$ and $\hat{\beta}_1$, which are equivalently characterised by each of the following:
  - $\hat{\beta}_0, \hat{\beta}_1 := \arg\min_{b_0,b_1} \hat{\mathbb{E}}(Y - b_0 - b_1 X)^2$,
  - $\hat{\beta}_0 = \hat{\mathbb{E}}Y - \hat{\beta}_1 \hat{\mathbb{E}}X, \hat{\beta}_1 = \frac{\hat{cov}(X,Y)}{\hat{var}(X)}$, and
  - $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{u}$, where $\hat{\mathbb{E}}\hat{u} = \hat{\mathbb{E}}X\hat{u} = 0$.
- Given that $\hat{\mathbb{E}}, \hat{cov}, \hat{var}$ are consistent estimators of their population counterparts, $\hat{\beta}_0, \hat{\beta}_1$ are consistent estimators of the population regression parameters. If orthogonality holds, then the population regression parameters coincide with the causal model parameters, and $\hat{\beta}_0, \hat{\beta}_1$ are also consistent estimators of the causal model parameters.
- $\hat{u}_i$ is a consistent estimator of e_i for $i \in \{1, \ldots, n\}$.
- The sample linear regression function is $y = \hat{\beta}_0 + \hat{\beta}_1 x$, the predicted values are $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ for $i \in \{1, \ldots, n\}$.
- Sample linear regression is also referred to as ordinary least squares regression.

Multiple Linear Regression

Causal Model

- For multiple linear regression with two variables, the causal model simplifies to $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$, where $u := \beta_3 x_3 + \ldots$ collects the remaining terms in the causal model of y.
- What follows straightforwardly extends to multiple linear regression with more than two variables.
- Note that β_0 and β_1 in this representation of the causal model and in the earlier representation of the causal model refer to the same parameter.

Population Regression

- The population linear regression problem, in this context, is the problem of constructing the best linear predictor of Y given X_1 and X_2. Formally, this problem is $\min_{b_0,b_1,b_2} \mathbb{E}(Y - b_0 - b_1 X_1 - b_2 X_2)$.
- The first-order conditions are:
  - $$\begin{aligned} FOC_{b_I} : \frac{\partial}{\partial b_1} C(b_0, b_1, b_2) &= \frac{\partial}{\partial b_I} \mathbb{E}(Y - b_0 - b_1 X_1 - b_2 X_2)^2 \\ &= \mathbb{E}[\frac{\partial}{\partial b_I}(Y - b_0 - b_1 X_1 - b_2 X_2)^2] \\ &= \mathbb{E}[2(Y - b_0 - b_1 X_1 - b_2 X_2)(-X_I)] \\ &= 0 \\ \Rightarrow \mathbb{E}[X_I(Y - b_0 - b_1 X_1 - b_2 X_2)] &= 0 \end{aligned},$$
  - for $I \in \{0, 1, 2\}$, where $X_0 := 1$.
  - Similarly, denote the solution to this problem as ρ_0, ρ_1, ρ_2.
- Then, as in simple linear regression, $Y = \rho_0 + \rho_1 X_1 + \rho_2 X_2 + e$ is the unique decomposition of Y into a linear function of X_1, X_2, with an error term e that has mean zero and is orthogonal to X_1, X_2.
- As in simple linear regression, supposing that orthogonality holds in the causal model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$, i.e. $\mathbb{E}u = \mathbb{E}X_1 u = \mathbb{E}X_2 u = 0$, the parameters of the causal parameter solve the population linear regression problem, and $\beta_0, \beta_1, \beta_2, u$ coincide with $\rho_0, \rho_1, \rho_2, e$.
- As in simple linear regression, supposing further that mean independence holds in the causal model, the population linear regression function coincides with the conditional expectation $\mathbb{E}[Y|X_1, X_2]$.
- As in simple linear regression, if orthogonality fails, the population linear regression model remains well-defined, and can be interpreted as a descriptive model that describes Y as an optimal linear function of X_1, X_2. Its parameters describe the average change in Y associated with some change in X_1 (X_2), holding X_2 (X_1) constant (supposing that the linear function constitutes a reasonable approximation of the conditional expectation).

Frisch-Waugh-Lovell Theorem

- Consider a population linear regression of X_1 on X_2, $X_1 = \pi_0 + \pi_2 X_2 + \tilde{X}_1$, where $\mathbb{E}\tilde{X}_1 = \mathbb{E}X_2\tilde{X}_1 = 0$ by construction. (The sense in which this result is by construction is explained below.) $\tilde{X}_1$ can be interpreted as the part of X_1 that is uncorrelated with X_2.
- Substitute the population linear regression of X_1 on X_2 into the causal model.
  - $$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \\ &= \beta_0 + \beta_1(\pi_0 + \pi_2 X_2 + \tilde{X}_1) + \beta_2 X_2 + u \\ &= (\beta_0 + \beta_1 \pi_0) + \beta_1 \tilde{X}_1 + (\beta_2 + \beta_1 \pi_2) X_2 + u \\ &=: \gamma_0 + \beta_1 \tilde{X}_1 + \gamma_2 X_2 + u \end{aligned}.$$

- Then, supposing that the causal model satisfies orthogonality,
  - $\begin{aligned} cov(Y, \tilde{X}_1) &= cov(\gamma_0 + \beta_1 \tilde{X}_1 + \gamma_2 X_2 + u, \tilde{X}_1) \\ &= cov(\gamma_0, \tilde{X}_1) + \beta_1 cov(\tilde{X}_1, \tilde{X}_1) + \gamma_2 cov(X_2, \tilde{X}_1) + cov(u, \tilde{X}_1) \\ &= \beta_1 var(\tilde{X}_1) \end{aligned}$, where
  - $cov(\gamma_0, \tilde{X}_1) = 0$ because γ_0 is a constant, $cov(X_2, \tilde{X}_1) = 0$ by construction of the population linear regression of X_1 on X_2, and $cov(u, \tilde{X}_1) = cov(u, X_1 - \pi_0 - \pi_2 X_2) = cov(u, X_1) - \pi_2 cov(u, X_2) = 0$ by orthogonality of u to X_1, X_2 in the causal model.
- Hence, supposing that the causal model satisfies orthogonality, $\beta_1 = \frac{cov(Y, \tilde{X}_1)}{var(\tilde{X}_1)}$.
- The Frisch-Waugh-Lovell theorem suggests a second method for computing the population regression coefficient, which consists in first computing the residual $\tilde{X}_1$ in the regression of X_1 on X_2 and then computing the regression coefficient of Y on $\tilde{X}_1$. The first method is to solve the population linear regression problem (of minimising the mean squared prediction error) "directly".
- Regression of Y on X_1 directly does not in general recover the causal parameter β_1 because this parameter is recovered only if $\beta_2 X_2 + u$ is orthogonal to X_1, which is the case if u is orthogonal to X_1, and $\beta_2 = 0$ or $cov(X_1, X_2) = 0$ (i.e. X_2 has no causal effect on Y, or X_1 and X_2 are uncorrelated). Neither of the latter two conditions are particularly "likely".

Sample Regression

- As in simple linear regression, the population parameters can be consistently estimated by the parameters $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ which solve the sample analogue of the population linear regression problem, namely $\min_{b_0, b_1, b_2} \hat{\mathbb{E}}(Y - b_0 - b_1 X_1 - b_2 X_2)$.
  - This is the construction (definition) of the sample regression parameters $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$.
- The first-order conditions of the sample linear regression problem are exactly analogous to that of the population linear regression problem, and have analogous implication:
  - $\hat{\mathbb{E}}[X_I(Y - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2)] = 0$ for $I \in \{0, 1, 2\}$, where $X_0 := 1$, where Y, X_1, X_2 in this instance refer to arbitrary observations in the sample, rather than arbitrary observations in the population.
- The sample linear regression residual is $\hat{u} := Y - \hat{\beta}_0 - \hat{\beta}_1 X_1 - \hat{\beta}_2 X_2$. Then, by construction of this residual and the sample linear regression parameters, $\hat{\mathbb{E}} X_I \hat{u} = 0$ for $I \in \{0, 1, 2\}$, where $X_0 = 1$, hence $c\hat{o}v(X_1 \hat{u}) = c\hat{o}v(X_2 \hat{u}) = 0$.
  - This result follows by construction of the residual and the parameters, and holds independent of whether the causal model satisfies orthogonality.
- As in simple linear regression, supposing that the causal model satisfies orthogonality, the sample linear regression parameters are consistent estimators of the parameters of the causal model (because the former are consistent estimators of the population linear regression parameters and if orthogonality holds, the population linear regression parameters coincide with the parameters of the causal model).
  - Otherwise, the population linear regression model is a (merely) descriptive model that describes Y as an optimal linear function of X_1, X_2, and the sample linear regression estimators are consistent estimators for the parameters of (only) the population linear regression model.
- The sample analogue of $\frac{cov(Y, \tilde{X}_1)}{var(\tilde{X}_1)}$ is a consistent estimator of that parameter. The sample linear regression of X_1 on X_2 (where in this context, these symbols denote arbitrary observations in the sample), is $X_1 = \hat{\pi} + \hat{\pi}_2 X_2 + \hat{\tilde{X}}_1$. Then, the estimator of the population linear regression parameter ρ_1 (and of the causal effect β_1 if the causal model satisfies orthogonality) is $\frac{c\hat{o}v(Y, \hat{\tilde{X}}_1)}{v\hat{a}r(\hat{\tilde{X}}_1}$.

Perfect Multicollinearity

- Some subset of regressors is perfectly multicollinear iff regression of one of these regressors on the rest yields a residual that is exactly zero. Informally, this is iff the regressors "perfectly explain" each other. For example, age and date of birth are perfectly multicollinear. The problem of perfect multicollinearity is resolved by dropping exactly one of the perfectly multicollinear variables.
- In the case where the perfectly multicollinear regressors are mutually exclusive and collectively exhaustive dummy variables, when one of these regressors is omitted, the coefficients on the remaining regressors should be interpreted as the difference in outcomes between the included and the omitted category.

Model Specification

- (See Duffy, 2022 - Notes on Linear Regression 2, pp. 45-49.) Any model that can be written in the form $Y = \beta_0 + \beta_1 W_1 + \ldots + u$ is linear in the parameters.