12.2. (a) When there is only one X, we only need to check that the instrument enters the first stage population regression. Since the instrument is $Z = X$, the regression of X onto Z will have a coefficient of 1.0 on Z, so that the instrument enters the first stage population regression. Key Concept 4.3 implies corr(X_i, u_i) = 0, and this implies corr(Z_i, u_i) = 0. Thus, the instrument is exogenous.

(b) Condition 1 is satisfied because there are no W's. Key Concept 4.3 implies that condition 2 is satisfied because (X_i, Z_i, Y_i) are i.i.d. draws from their joint distribution. Condition 3 is also satisfied by applying assumption 3 in Key Concept 4.3. Condition 4 is satisfied because of conclusion in part (a).

(c) The TSLS estimator is $\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}}$ using Equation (10.4) in the text. Since Z_i = X_i, we have

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} = \frac{s_{XY}}{s_X^2} = \hat{\beta}_1^{OLS}.$$

12.4.

$$\hat{\beta}_{TSLS} = \frac{s_{ZY}}{s_{ZX}} = \frac{\frac{1}{\mathrm{n}-1}\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\frac{1}{\mathrm{n}-1}\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})} = \frac{\sum_{i=1}^{n} Z_i(Y_i - \bar{Y})}{\sum_{i=1}^{n} Z_i(X_i - \bar{X})}$$

$$= \frac{\sum_{i=1}^{n} Z_i Y_i - \sum_{i=1}^{n} Z_i \overline{\mathrm{Y}}}{\sum_{i=1}^{n} Z_i X_i - \sum_{i=1}^{n} Z_i \overline{\mathrm{X}}} = \frac{n_1\left(\overline{Y_{|Z=1}} - \bar{Y}\right)}{n_1\left(\overline{X_{|Z=1}} - \bar{X}\right)} = \frac{\left(\overline{Y_{|Z=1}} - \frac{1}{n}\left(n_1\overline{Y_{|Z=1}} + n_0\overline{Y_{|Z=0}}\right)\right)}{\left(\overline{X_{|Z=1}} - \frac{1}{n}\left(n_1\overline{\mathrm{X}_{|Z=1}} + n_0\overline{\mathrm{X}_{|Z=0}}\right)\right)}$$

$$= \frac{\left(n\overline{Y_{|Z=1}} - \left(n_1\overline{Y_{|Z=1}} + n_0\overline{Y_{|Z=0}}\right)\right)}{\left(n\overline{X_{|Z=1}} - \left(n_1\overline{X_{|Z=1}} + n_0\overline{X_{|Z=0}}\right)\right)} = \frac{\left((n_1+n_0)\overline{Y_{|Z=1}} - \left(n_1\overline{Y_{|Z=1}} + n_0\overline{Y_{|Z=0}}\right)\right)}{\left((n_1+n_0)\overline{X_{|Z=1}} - \left(n_1\overline{X_{|Z=1}} + n_0\overline{X_{|Z=0}}\right)\right)}$$

$$= \frac{\overline{Y_{|Z=1}} - \overline{Y_{|Z=0}}}{\overline{X_{|Z=1}} - \overline{X_{|Z=0}}}$$

Where

$\sum_{i=1}^{n} Z_i\overline{\mathrm{Y}} = \overline{\mathrm{Y}}\sum_{i=1}^{n} Z_i = \overline{\mathrm{Y}}\cdot(0+1+1+0+\cdots+0+1) = \overline{\mathrm{Y}}\,n_1,$

$\sum_{i=1}^{n} Z_i Y_i = (0\cdot Y_1 + 1\cdot Y_2 + 1\cdot Y_3 + 0\cdot Y_4 + \cdots + 0\cdot Y_{n-1} + 1\cdot Y_n) = n_1\overline{Y_{|Z=1}},$

$\overline{Y} = \frac{1}{n}\left(n_1\overline{Y_{|Z=1}} + n_0\overline{Y_{|Z=0}}\right)$ and

$$\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y}) = \sum_{i=1}^{n} Z_i(Y_i - \bar{Y}) - \sum_{i=1}^{n}\bar{Z}(Y_i - \bar{Y})$$

$$= \sum_{i=1}^{n} Z_i(Y_i - \bar{Y}) - \sum_{i=1}^{n}\bar{Z}Y_i + \sum_{i=1}^{n}\bar{Z}\bar{Y} = \sum_{i=1}^{n} Z_i(Y_i - \bar{Y}) - n\bar{Z}\bar{Y} + n\bar{Z}\bar{Y}$$

$$= \sum_{i=1}^{n} Z_i(Y_i - \bar{Y})$$

Panel B: Wald Estimates for 1980 Census—Men Born 1930–1939

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Difference (std. error) (1) − (2)
ln (wkly. wage)	5.8916	5.9027	−0.01110
			(0.00274)
Education	12.6881	12.7969	−0.1088
			(0.0132)
Wald est. of return to education			0.1020
			(0.0239)
OLS return to education			0.0709
			(0.0003)

a. The sample size is 247,199 in Panel A, and 327,509 in Panel B. Each sample consists of males born in the United States who had positive earnings in the year preceding the survey. The 1980 Census sample is drawn from the 5 percent sample, and the 1970 Census sample is from the State, County, and Neighborhoods 1 percent samples.

b. The OLS return to education was estimated from a bivariate regression of log weekly earnings on years of education.

12.6. Using the homoskedasitic-only *F*-statistic, which must be greater than 10 to be considered a strong instrument, we have

$$F_{HomoskedasitcOnly} = \frac{\frac{R^2 - R_r^2}{q}}{\frac{(1-R^2)}{N-k-1}} = \frac{\frac{R^2-0}{1}}{\frac{(1-R^2)}{(113-1-1)}} = \frac{R^2}{\frac{(1-R^2)}{111}} > 10$$

$$\frac{R^2}{1-R^2} > \frac{10}{111} \Leftrightarrow \frac{1-R^2}{R^2} < \frac{111}{10} \Leftrightarrow \frac{1}{R^2} - 1 < \frac{111}{10} \Leftrightarrow \frac{1}{R^2} < \frac{111}{10} + \frac{10}{10}$$

$$\Leftrightarrow \frac{1}{R^2} < \frac{121}{10} \Leftrightarrow R^2 > \frac{10}{121} \Leftrightarrow R < -\frac{\sqrt{10}}{11} \vee R > \frac{\sqrt{10}}{11}$$

Where $R^2 = r^2$ with *r* as the estimated correlated coefficient between X_i and Z_i.

Thus, the range where *r* satisfies the inequality can be found to be $r > \frac{+\sqrt{10}}{11}$ and $r < \frac{-\sqrt{10}}{11}$ where $\frac{+\sqrt{10}}{11} \approx 0.29$.

12.8. (a) $Q_i^s = \beta_0 + \beta_1 P_i + u_i^s$ (Main Equation)

$Q_i^d = \gamma_0 + u_i^d$

$Q_i^s = Q_i^d$

Equilibrium price: $P = \frac{\gamma_0 - \beta_0}{\beta_1} + \frac{u_i^d - u_i^s}{\beta_1}$ (First Stage)

So $cov(P,\ u^s) = cov\left(\frac{\gamma_0 - \beta_0}{\beta_1} + \frac{u_i^d - u_i^s}{\beta_1},\ u_i^s\right) = cov\left(\frac{\gamma_0 - \beta_0}{\beta_1},\ u_i^s\right) + cov\left(\frac{u_i^d - u_i^s}{\beta_1},\ u_i^s\right)$

$= \frac{1}{\beta_1}\{cov(u_i^d,\ u_i^s) - cov(u_i^s,\ u_i^s)\} = \frac{-var(u_i^s)}{\beta_1} = \frac{-\sigma_{u^s}^2}{\beta_1} \neq 0$

(b) Because $\hat{\beta}_1^{OLS} = \frac{s_{PQ}}{s_P^2} \xrightarrow{p} \frac{cov(P,\ Q)}{var(P)} = \beta_1 + \frac{cov(P,\ u^s)}{var(P)}$, and $cov(P,\ u^s) \neq 0$, the OLS estimator (where we regress Q on P) is inconsistent for β_1.

(c) We need an instrumental variable, something that is correlated with P but uncorrelated with u^s. The variation in P is driven by $u_i^d - u_i^s$, so that suggests that we would need an instrument Z that is part of $u_i^d = \alpha_0 + \alpha_1 Z_i + \tilde{u}_i^d$. That is, we need a shock in demand that does not relate to supply shocks u_i^s.

In this case Q can serve as the instrument, because demand is completely inelastic *and* demand shocks do not correlate to supply shocks ($cov(u_i^d,\ u_i^s) = 0$). This means Q is exogenous ($cov(Q,\ u_i^s) = cov(\gamma_0 + u_i^d,\ u_i^s) = 0$) and relevant for P at the same time ($cov(Q,\ P) = \frac{var(u_i^d)}{\beta_1} \neq 0$). The γ_0 can be estimated by OLS (equivalently as the sample mean of Q_i).

Clearly this case is weird, using Y as an instrument for X. In practice we will need an exogenous demand shock to estimate supply (of umbrellas and weather), or an exogenous supply shock to estimate demand (for fish and weather at sea).

12.9. (a) While it is plausible that having more secure property rights affects a country's level of income, it is also plausible that richer countries can afford institutions that protect property rights. Therefore, there is a problem of simultaneous causality bias. According to this theory, as richer countries can afford secure property rights, OLS estimates might be overestimated.

(b) Settler mortality is exogenous if it is not directly related to a country's recent income or is not related to unobserved factors affecting both, recent protection against expropriation and income. It is relevant if it affected European settlement, and subsequently led to the establishment of early institutions for securing property rights, which have persisted until the present day. Taking these two together, the instrument is valid if settler mortality is a strong predictor of recent protection against expropriation, and if settler mortality is related to the other variables in the model in the way set out by the theory around the causal pathway. Note that in this model, because the instrument is not purely random, the use of included exogenous regressors and control variables is important in order for the instrument exogeneity assumption to be plausibly valid.

13.11. Following the notation used in Chapter 13, let π_{1i} denote the coefficient on distance to the nearest hospital offering cardiac catheterization in the "first stage" IV regression, and let β_{1i} denote the effect of cardiac catheterization on patient survival times. From Equation (13.12)

$$\begin{aligned}\hat{\beta}^{TSLS} \xrightarrow{p} \frac{E(\beta_{1i}\ \pi_{1i})}{E(\pi_{1i})} &= \frac{E(\beta_{1i})E(\pi_{1i}) + cov(\beta_{1i}, \pi_{1i})}{E(\pi_{1i})} \\ &= E(\beta_{1i}) + \frac{cov(\beta_{1i}, \pi_{1i})}{E(\pi_{1i})} \\ &= \text{Average treatment effect} + \frac{cov(\beta_{1i}, \pi_{1i})}{E(\pi_{1i})}\end{aligned}$$

where the first equality uses the properties of covariances (equation (2.34)), and the second equality uses the definition of the average treatment effect. Evidently, the local average treatment effect will deviate from the average treatment effect when $cov(\beta_{1i}, \pi_{1i}) \neq 0$. As discussed in Section 13.6, this covariance is zero when β_{1i} or π_{1i} are constant. This seems likely. But, for the sake of argument, suppose that they are not constant; that is, suppose the impact of catheterization on survival times varies across patients (β_{1i} is not constant) as does the effect of distances on the use of catheterization (π_{1i} is not constant). Are β_{1i} and π_{1i} related? They might be if we believe that the estimated effect is different for people who are sensitive to distance from a hospital. For instance, very old individuals may be more sensitive to distance from a hospital for getting treatment. They might also be more or less sensitive to the treatment compared to other types of patients. For example, let us assume that older people benefit more from the use of catheterization. This suggests that β_{1i} and π_{1i} are positively related, so that $cov(\beta_{1i},\pi_{1i}) > 0$. Because $E(\pi_{1i}) < 0$, this suggests that the local average treatment effect is lesser than the average treatment effect when β_{1i} varies between older and younger patients.