Supplementary Appendix

Humans in the Loop: The Next Frontier in the Credibility Revolution

Megan T. Stevenson and Joshua B. Fischman

July 25, 2026

A Appendix

A.1 Simulation details

To simulate the linear probability model, we use the same two-stage specification as in Equations (1) and (2), displayed below for convenience:

$$D = c_1 + \delta Z + \gamma_1 X + v \tag{1}$$
$$Y = c_2 + \beta D + \gamma_2 X + u \tag{2}$$

In the first set of simulations, the outcome Y, treatment D, instrument Z, and unobserved characteristic X are all dichotomous. For simplicity, $\gamma_1 = \gamma_2 = \gamma$, and there are no constant terms in these equations. We define ρ as $corr(\gamma X + v, \gamma X + u)$ and interpret it as the degree of endogeneity. In this setting, endogeneity is driven solely by the strength of the omitted variable X and is implemented numerically by adjusting γ. v and u are random noise, uncorrelated with X, Z or each other.

We use the bisection method to identify values of δ that would result in simulated values of $E[F] = 20, 80, 200$. For each such value of δ, we then use bisection to choose γ to achieve the specified ρ. To achieve our target range of $E[F]$ and ρ and ensure that the predicted probabilities are between 0 and 1, we set $P(Z = 1) = 0.5$ and $P(X = 1) = 0.7$. We simulate D and Y according to the data generating process: $P(D = 1) = \delta Z + \gamma_1 X$ and $P(Y = 1) = \beta D + \gamma_2 X$. We simulated samples of 1000 observations, iterated 200,000 times.

For the continuous model, we use the same specification as above, but with all independent variables normally distributed. We chose $c_1 = c_2 = 0$ and $u, v, X, Z \sim N(0, 1)$. We set $\gamma_1 = \gamma_2 = \sqrt{\frac{\rho}{1-\rho}}$. We then used bisection to find values of δ that would achieve $E[F] = 20, 80, 200$.

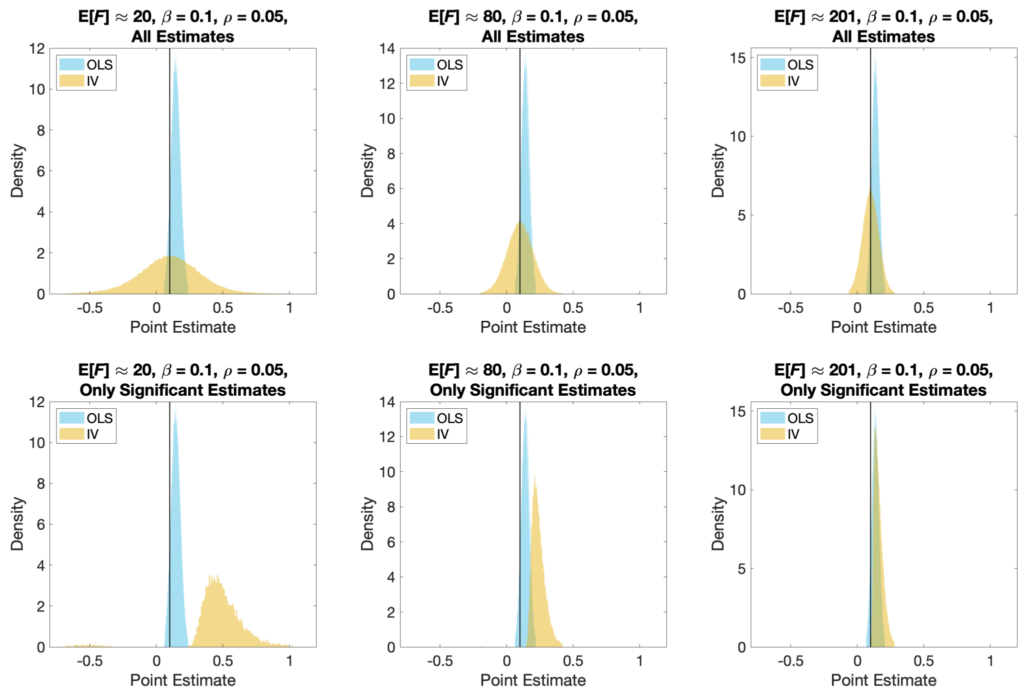
A.2 Simulations

The following simulations vary ρ, β and $E[F]$ within both a linear probability model and a continuous model. Details of the simulations can be found in Appendix A.1. The figure titles indicate the type of model and the subfigure headings show the parameters chosen. The figures alternate between $\beta = 0.1$ and $\beta = 0$. The first two figures show relatively low endogeneity, $\rho = 0.05$. The next two figures show $\rho = 0.15$ and then $\rho = 0.25$. We show two figures for the continuous model, both with $\rho = 0.15$.

Figures A.1-A.8 show that, under significance bias, IV is frequently more biased than OLS. The relative bias depends in part on the degree of endogeneity; when treatment is only mildly endogenous, as in Figures A.1 and A.2, OLS is more likely to outperform IV. The degree of bias also depends on the true value of β. When β is large relative to the standard error, as shown in the bottom right corner of Figures A.1, A.3, A.5, and A.7, results are more likely to be statistically significant and significance bias has less impact. When β=0, significance bias cuts out a large chunk from the center of the distribution. IV is double-humped and can be heavily biased, as shown in Figures A.1, A.2, A.4, A.6, and A.8. In all figures, the degree of bias among IV estimates decreases as the strength of the first-stage F statistic increases.

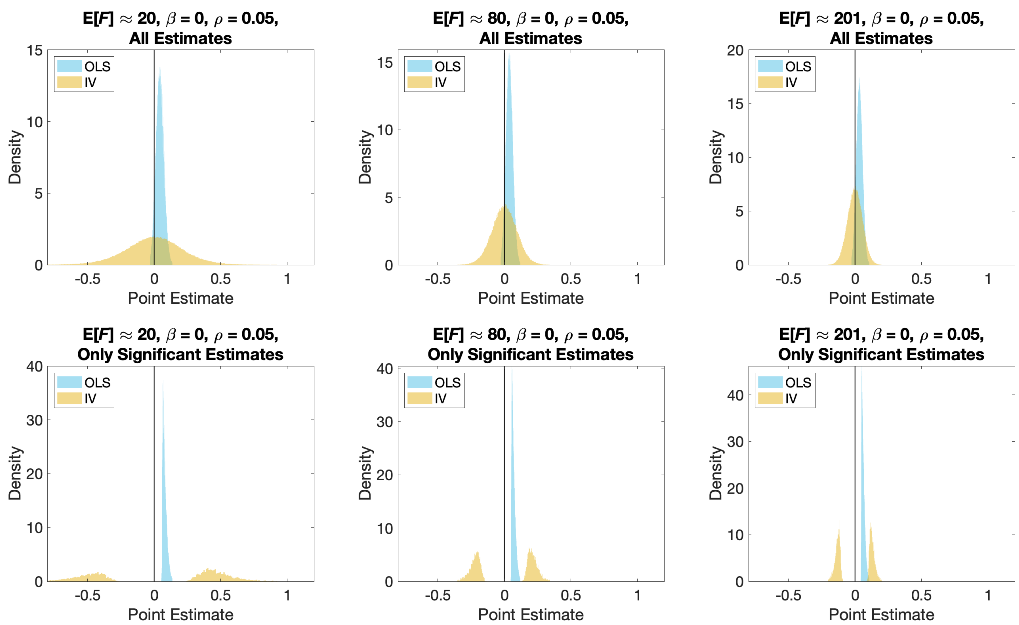
Figures A.9-A.12 show how the degree of bias varies with the threshold under significance bias and threshold testing. Figures A.10-A.11 show that qualitative patterns for the continuous model are very similar to those shown in the main text for the linear probability model when it comes to median bias and median absolute bias: bias decreases with the threshold. Figures A.9 and A.12 show the mean bias for the linear probability and continuous models, respectively. The mean bias behaves similarly to the median bias when $\beta = 0.1$ but increases with the threshold when $\beta = 0$. However, the mean does not exist in the IV estimate, and may be particularly unstable when instruments are weak. It is therefore not a preferred measure of bias.

Figure A.1: Distributions of estimators with and without significance bias: linear probability model, $\beta = 0.1, \rho = 0.05$



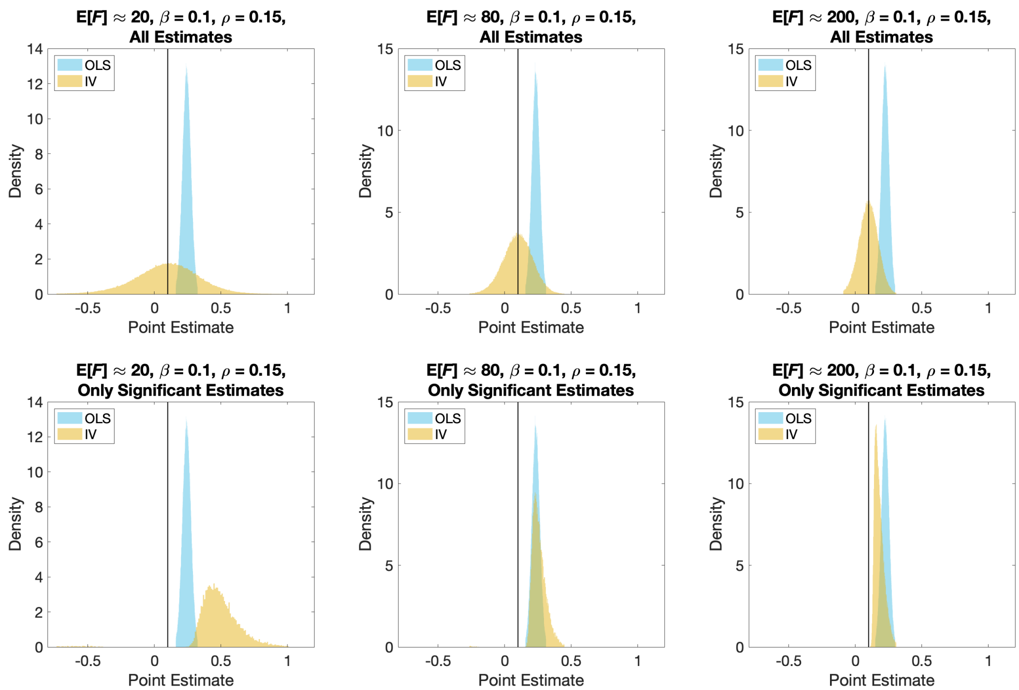
Note: The top row shows distributions of the IV and OLS estimators without significance bias, using the parameters shown. The second row shows the distributions of the IV and OLS estimators conditional on statistical significance. Each simulated sample size is 1000 and there are 200,000 simulations.

Figure A.2: Distributions of estimators with and without significance bias: linear probability model, $\beta = 0, \rho = 0.05$



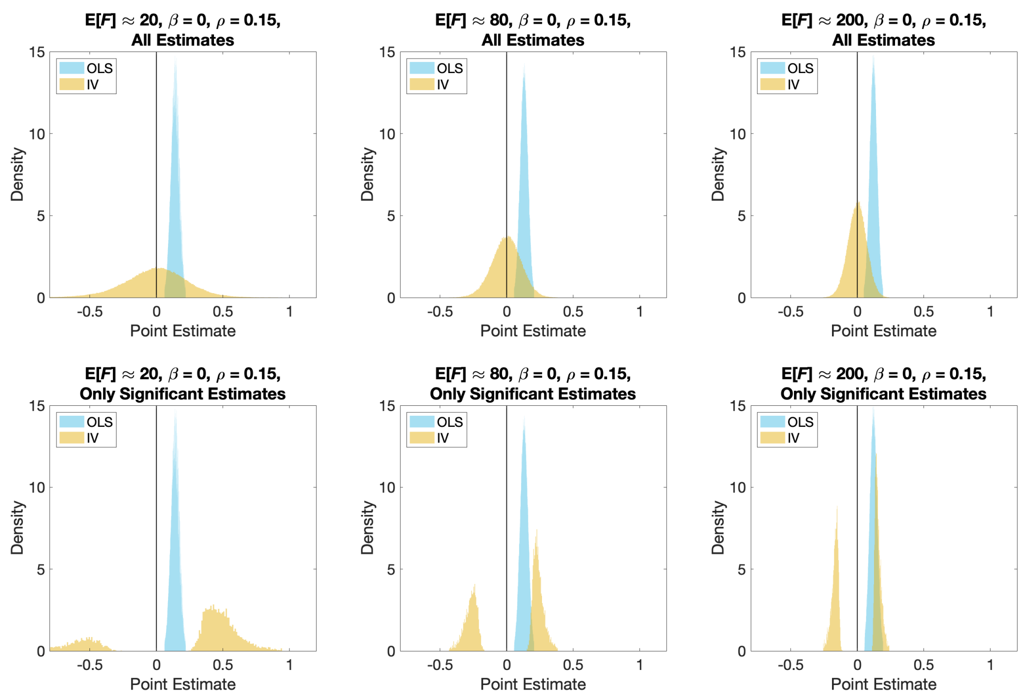
Note: The top row shows distributions of the IV and OLS estimators without significance bias, using the parameters shown. The second row shows the distributions of the IV and OLS estimators conditional on statistical significance. Each simulated sample size is 1000 and there are 200,000 simulations.

Figure A.3: Distributions of IV and OLS estimators with and without significance bias, linear probability model, $\beta = 0.1, \rho = 0.15$



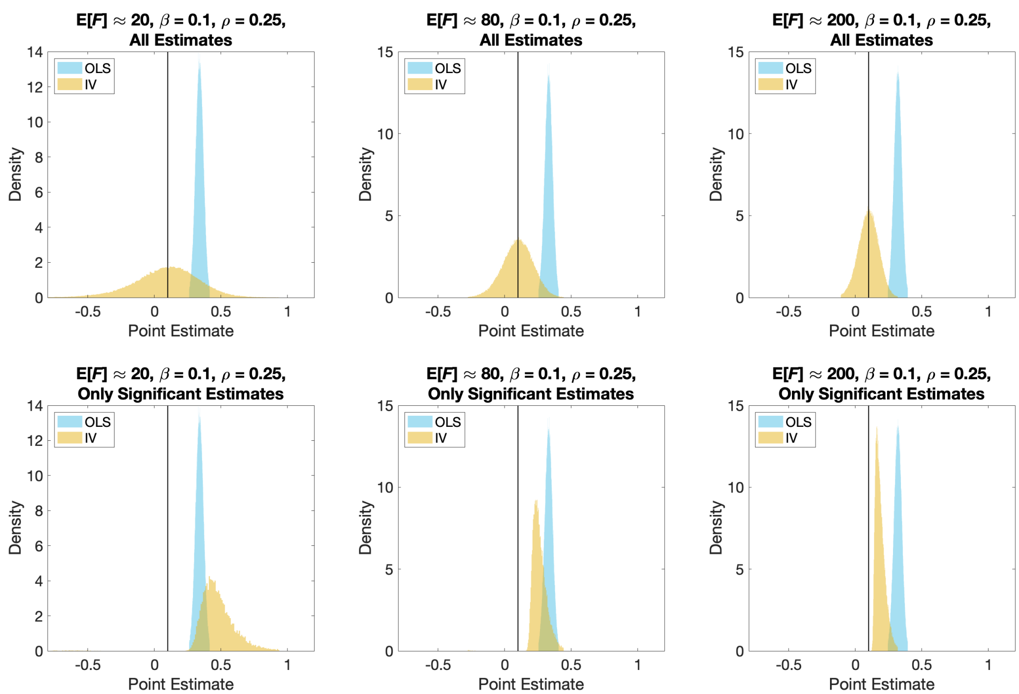
Note: The top row shows distributions of the IV and OLS estimators without significance bias, using the parameters shown. The second row shows the distributions of the IV and OLS estimators conditional on statistical significance. Each simulated sample size is 1000 and there are 200,000 simulations.

Figure A.4: Distributions of estimators with and without significance bias: linear probability model, $\beta = 0, \rho = 0.15$



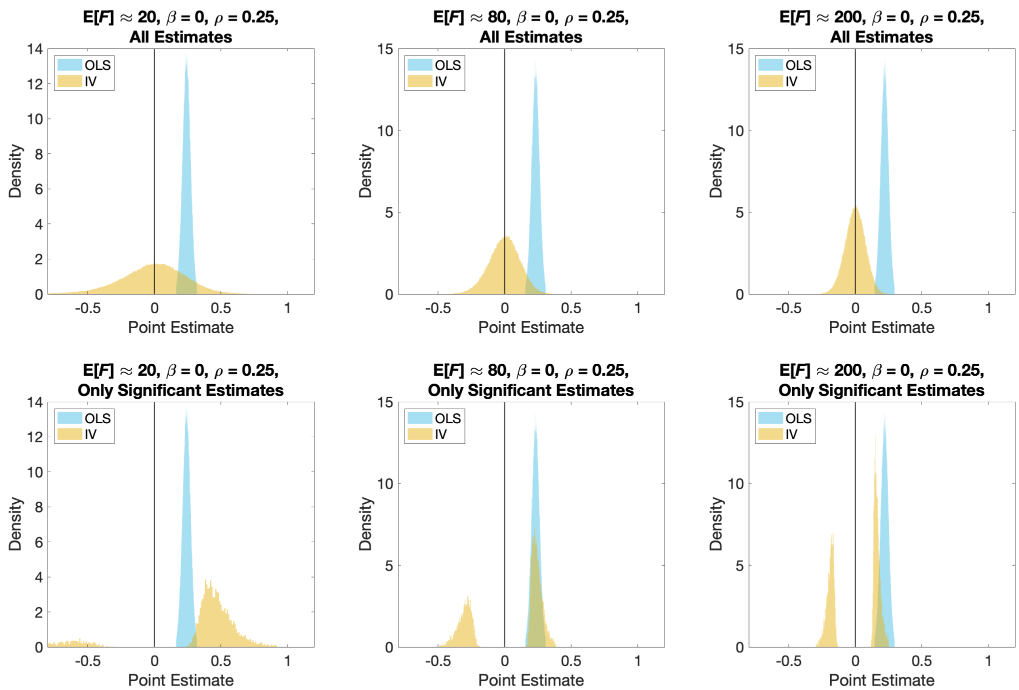
Note: The top row shows distributions of the IV and OLS estimators without significance bias, using the parameters shown. The second row shows the distributions of the IV and OLS estimators conditional on statistical significance. Each simulated sample size is 1000 and there are 200,000 simulations.

Figure A.5: Distributions of estimators with and without significance bias: linear probability model, $\beta = 0.1, \rho = 0.25$



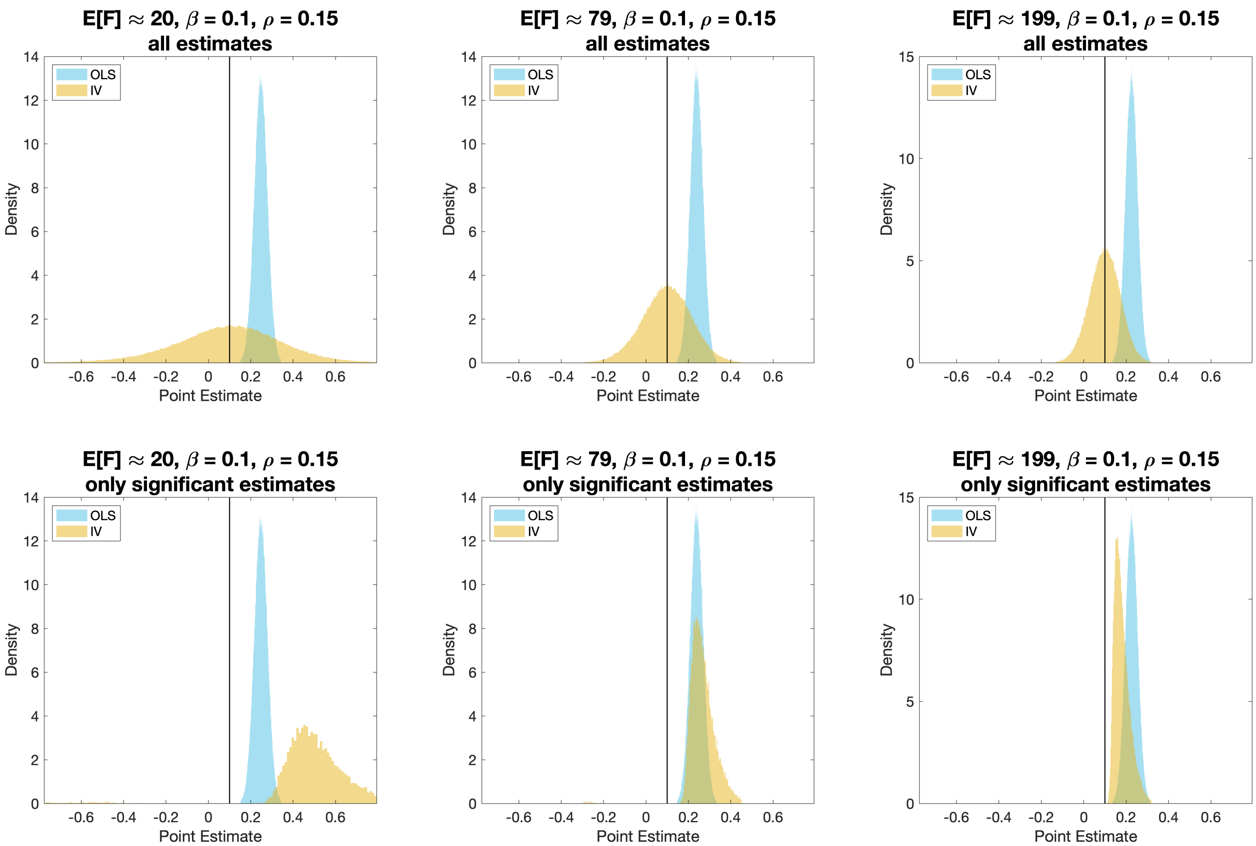
Note: The top row shows distributions of the IV and OLS estimators without significance bias, using the parameters shown. The second row shows the distributions of the IV and OLS estimators conditional on statistical significance. Each simulated sample size is 1000 and there are 200,000 simulations.

Figure A.6: Distributions of estimators with and without significance bias: linear probability model, $\beta = 0, \rho = 0.25$



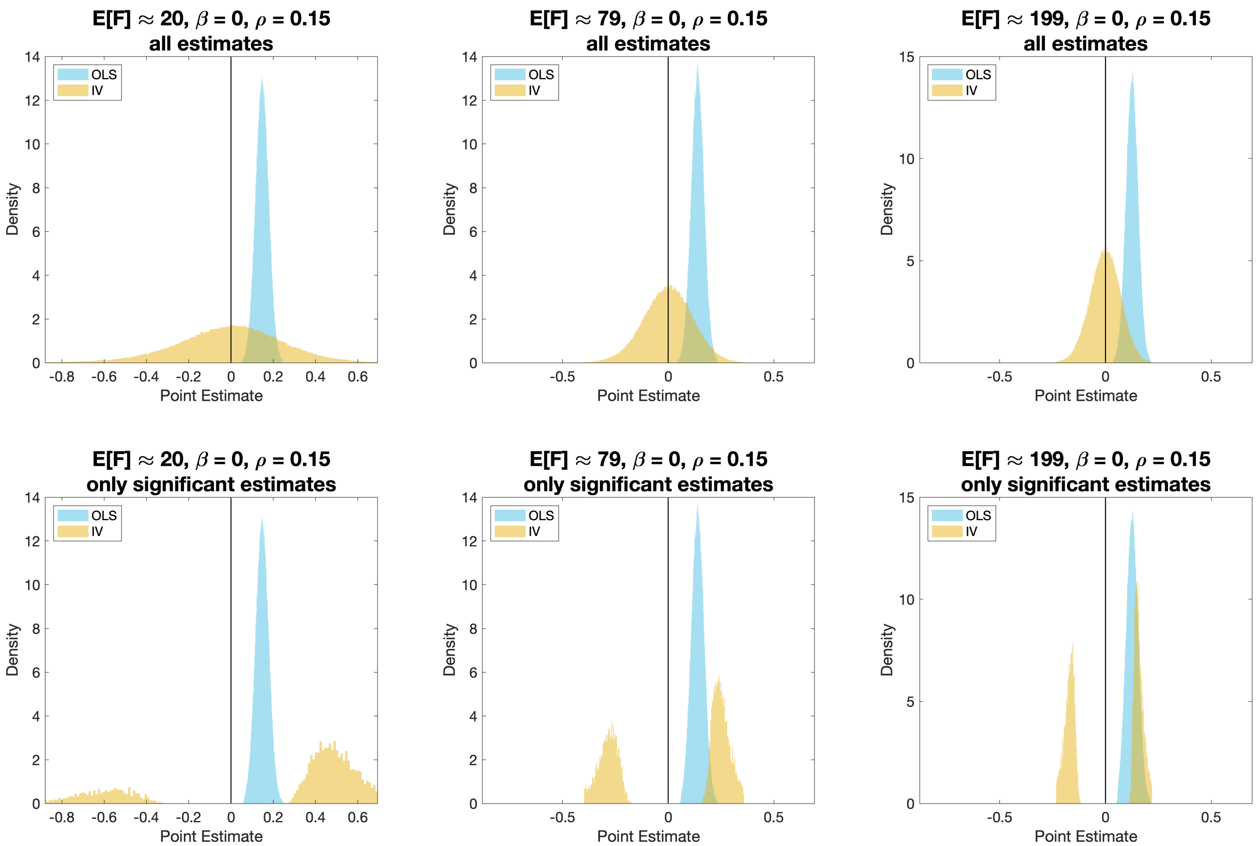
Note: The top row shows distributions of the IV and OLS estimators without significance bias, using the parameters shown. The second row shows the distributions of the IV and OLS estimators conditional on statistical significance. Each simulated sample size is 1000 and there are 200,000 simulations.

Figure A.7: Distributions of estimators with and without significance bias: continuous model, $\beta = 0.1, \rho = 0.15$



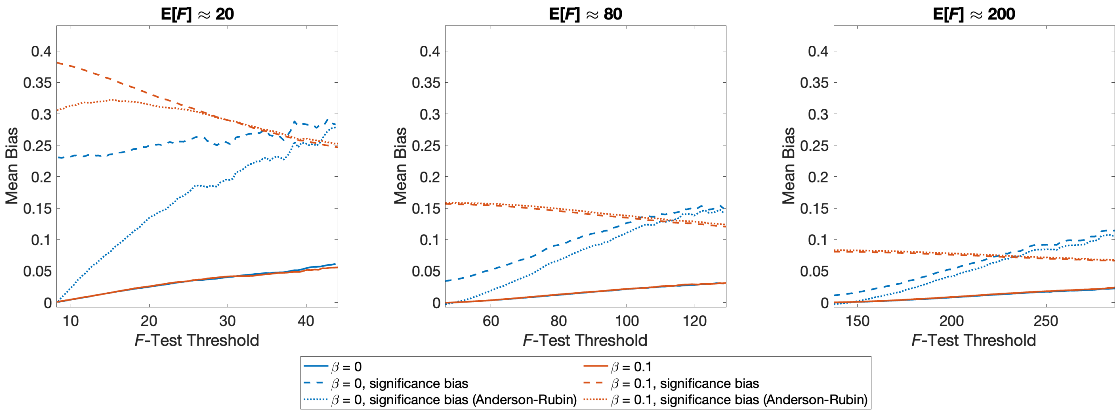
Note: The top row shows distributions of the IV and OLS estimators without significance bias, using the parameters shown. The second row shows the distributions of the IV and OLS estimators conditional on statistical significance. Each simulated sample size is 1000 and there are 200,000 simulations. The standard deviation of both D and Y is approximately 1.1.

Figure A.8: Distributions of estimators with and without significance bias: continuous model, $\beta = 0, \rho = 0.15$



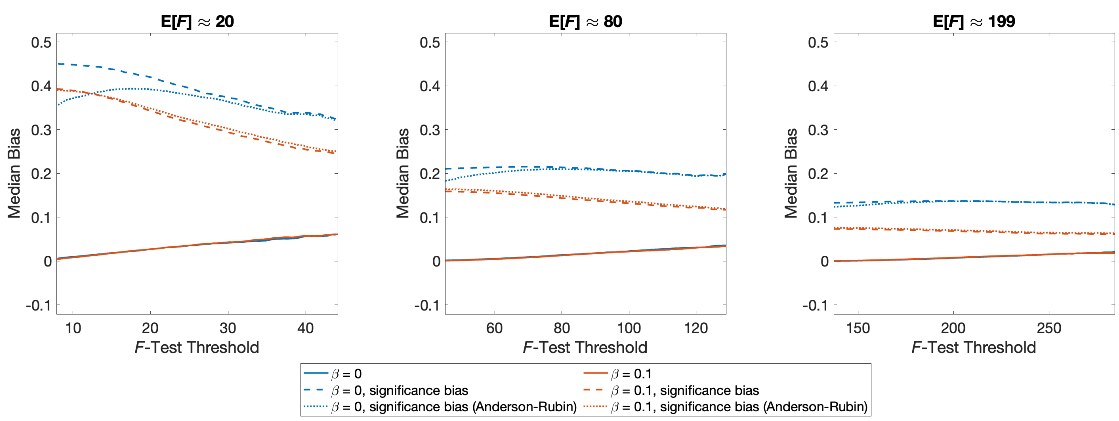
Note: The top row shows distributions of the IV and OLS estimators without significance bias, using the parameters shown. The second row shows the distributions of the IV and OLS estimators conditional on statistical significance. Each simulated sample size is 1000 and there are 200,000 simulations. The standard deviation of both D and Y is approximately 1.1.

Figure A.9: Mean bias as a function of the threshold test: linear probability model



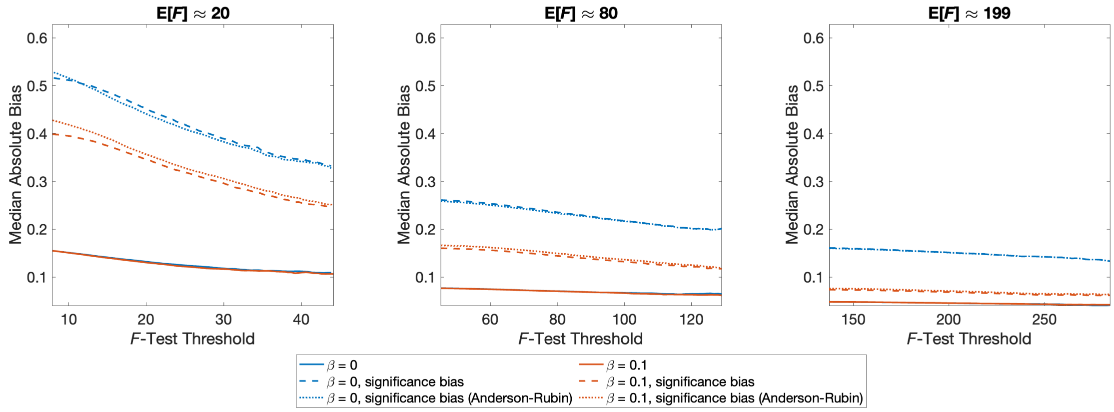
Note: The vertical axis shows the mean bias, where bias is defined as either $(\beta_{IV} - \beta)$ or $(\beta_{IV}^{SB} - \beta)$, among the set of simulated samples whose observed first-stage *F-statistic* is greater than the threshold shown on the horizontal axis. The simulations have binary outcome, treatment, and instrument with $\rho = 0.15$; more details can be found in Subsection 2.2 and Appendix A.1. Each of the 200,000 simulated samples has 1000 observations. The solid lines show all simulations, the dashed lines show all simulations that are statistically significant using heteroskedastic-robust standard errors and the dotted lines show all simulations that are statistically significant according to the Anderson-Rubin method.

Figure A.10: Median bias as a function of the threshold test: continuous model



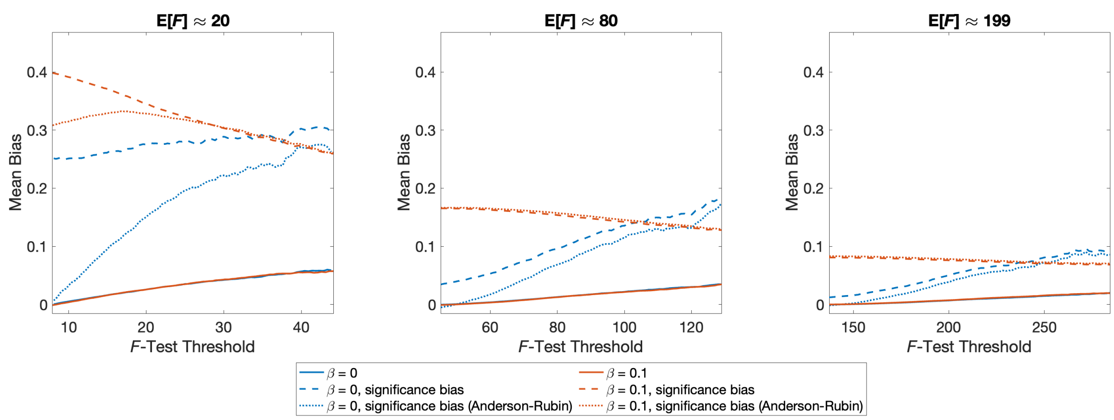
Note: The vertical axis shows the median bias, where bias is defined as either $(\beta_{IV} - \beta)$ or $(\beta_{IV}^{SB} - \beta)$, among the set of simulated samples whose observed first-stage *F-statistic* is greater than the threshold shown on the horizontal axis. The simulations have continuous outcome, treatment, and instrument with $\rho = 0.15$; more details can be found in Subsection 2.2 and Appendix A.1. The standard deviation of both D and Y is approximately 1.1. Each of the 200,000 simulated samples has 1000 observations. The solid lines show all simulations, the dashed lines show all simulations that are statistically significant using heteroskedastic-robust standard errors and the dotted lines show all simulations that are statistically significant according to the Anderson-Rubin method.

Figure A.11: Median absolute bias as a function of the threshold test: continuous model



Note: The vertical axis shows the median absolute bias, where bias is defined as either $(\beta_{IV} - \beta)$ or $(\beta_{IV}^{SB} - \beta)$, among the set of simulated samples whose observed first-stage *F-statistic* is greater than the threshold shown on the horizontal axis. The simulations have continuous outcome, treatment, and instrument with $\rho = 0.15$; more details can be found in Subsection 2.2 and Appendix A.1. The standard deviation of both D and Y is approximately 1.1. Each of the 200,000 simulated samples has 1000 observations. The solid lines show all simulations, the dashed lines show all simulations that are statistically significant using heteroskedastic-robust standard errors and the dotted lines show all simulations that are statistically significant according to the Anderson-Rubin method.

Figure A.12: Mean bias as a function of the threshold test: continuous model



Note: The vertical axis shows the mean bias, where bias is defined as either $(\beta_{IV} - \beta)$ or $(\beta_{IV}^{SB} - \beta)$, among the set of simulated samples whose observed first-stage *F-statistic* is greater than the threshold shown on the horizontal axis. The simulations have continuous outcome, treatment, and instrument with $\rho = 0.15$; more details can be found in Subsection 2.2 and Appendix A.1. The standard deviation of both D and Y is approximately 1.1. Each of the 200,000 simulated samples has 1000 observations. The solid lines show all simulations, the dashed lines show all simulations that are statistically significant using heteroskedastic-robust standard errors and the dotted lines show all simulations that are statistically significant according to the Anderson-Rubin method.