PREDICTION

True model $$P_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

Fitted model $$\hat{P}_i = \hat{\beta}_1 + \sum_{j=2}^{k} \hat{\beta}_j X_{ji}$$

В предыдущей последовательности мы увидели как прогнозировать цены на товары или активы, учитывая их структурные характеристики.(построение). В этом ряду мы обсудим свойства таких прогнозов.

PREDICTION

True model $$P_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

Fitted model $$\hat{P}_i = \hat{\beta}_1 + \sum_{j=2}^{k} \hat{\beta}_j X_{ji}$$

Предположим, что даны примеры n наблюдений, как показано, мы установили модель ценообразования с k-1 характеристиками.

PREDICTION

True model

$$P_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

Fitted model

$$\hat{P}_i = \hat{\beta}_1 + \sum_{j=2}^{k} \hat{\beta}_j X_{ji}$$

Prediction conditional on $\left\{X_2^*, X_3^*, \ldots, X_k^*\right\}$

$$\hat{P}^* = \hat{\beta}_1 + \sum_{j=2}^{k} \hat{\beta}_j X_j^*$$

Предположим сейчас, что оно сталкивается с новым видом товара с характеристиками $\{X_2^*, X_3^*, ..., X_k^*\}$. Учитывая (пример) результат выборки регрессии, естественно прогнозировать, что цена нового вида должна сводиться к третьему уравнению.

PREDICTION

True model

$$P_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

Fitted model

$$\hat{P}_i = \hat{\beta}_1 + \sum_{j=2}^{k} \hat{\beta}_j X_{ji}$$

Prediction conditional on $\left\{X_2^*, X_3^*, \ldots, X_k^*\right\}$

$$\hat{P}^* = \hat{\beta}_1 + \sum_{j=2}^{k} \hat{\beta}_j X_j^*$$

Что же можно сказать о свойствах этого прогнозирования? Во-первых, естественно спросить справедливо ли оно в НЕсистематической переоценке или недооценке смысла фактической цены. Во-вторых, мы обеспокоены о вероятной точности прогноза

PREDICTION

True model $P_i = \beta_1 + \beta_2 X_i + u_i$

Fitted model $\hat{P}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Prediction conditional on X^* $\hat{P}^* = \hat{\beta}_1 + \hat{\beta}_2 X^*$

Мы будем рассматривать пример, где товар имеет одну соответствующую характеристику и предположим, что мы установили простую регрессионную показанную модель. Следовательно, дадим новый вид товара с характеристиками $X = X^*$, эта модель дает нам прогнозируемую цену.

PREDICTION

True model $P_i = \beta_1 + \beta_2 X_i + u_i$

Fitted model $\hat{P}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Prediction conditional on X^* $\hat{P}^* = \hat{\beta}_1 + \hat{\beta}_2 X^*$

Actual value of P^* $P^* = \beta_1 + \beta_2 X^* + u^*$

Допустим, что модель относится (обращается) к новому товару и, следовательно, к фактической цене, условно $X = X^*$; оно образуется как показано, где u^* - это значение нарушения условия нового товара.

PREDICTION

True model $P_i = \beta_1 + \beta_2 X_i + u_i$

Fitted model $\hat{P}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Prediction conditional on X^* $\hat{P}^* = \hat{\beta}_1 + \hat{\beta}_2 X^*$

Actual value of P^* $P^* = \beta_1 + \beta_2 X^* + u^*$

$$PE = P^* - \hat{P}^*$$

Мы объясним ошибку прогнозирования модели PE, как разницу между ценой фактической и прогнозируемой

PREDICTION

True model $P_i = \beta_1 + \beta_2 X_i + u_i$

Fitted model $\hat{P}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Prediction conditional on X^* $\hat{P}^* = \hat{\beta}_1 + \hat{\beta}_2 X^*$

Actual value of P^* $P^* = \beta_1 + \beta_2 X^* + u^*$

$$PE = P^* - \hat{P}^* = \left(\beta_1 + \beta_2 X^* + u^*\right) - \left(\hat{\beta}_1 + \hat{\beta}_2 X^*\right)$$

Прогнозируемая ошибка получается путем подстановки фактической и прогнозируемой цены.

PREDICTION

True model $P_i = \beta_1 + \beta_2 X_i + u_i$

Fitted model $\hat{P}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Prediction conditional on X^* $\hat{P}^* = \hat{\beta}_1 + \hat{\beta}_2 X^*$

Actual value of P^* $P^* = \beta_1 + \beta_2 X^* + u^*$

$$PE = P^* - \hat{P}^* = \left(\beta_1 + \beta_2 X^* + u^*\right) - \left(\hat{\beta}_1 + \hat{\beta}_2 X^*\right)$$

$$E\left(PE\right) = E\left(\beta_1 + \beta_2 X^* + u^*\right) - E\left(\hat{\beta}_1 + \hat{\beta}_2 X^*\right)$$

Мы ожидаем.

PREDICTION

True model $P_i = \beta_1 + \beta_2 X_i + u_i$

Fitted model $\hat{P}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Prediction conditional on X^* $\hat{P}^* = \hat{\beta}_1 + \hat{\beta}_2 X^*$

Actual value of P^* $P^* = \beta_1 + \beta_2 X^* + u^*$

$$PE = P^* - \hat{P}^* = \left(\beta_1 + \beta_2 X^* + u^*\right) - \left(\hat{\beta}_1 + \hat{\beta}_2 X^*\right)$$

$$E(PE) = E\left(\beta_1 + \beta_2 X^* + u^*\right) - E\left(\hat{\beta}_1 + \hat{\beta}_2 X^*\right)$$

$$= \beta_1 + \beta_2 X^* + E\left(u^*\right) - E\left(\hat{\beta}_1\right) - X^* E\left(\hat{\beta}_2\right)$$

β_1 и β_2 приняты с фиксированными параметрами, поэтому ожидания на них не влияют. Также X^* принято считать фиксированной величиной и также ожидания не воздействуют. Тем не менее, u^* $\hat{\beta}_1$ и $\hat{\beta}_2$ случайные переменные.

PREDICTION

True model $P_i = \beta_1 + \beta_2 X_i + u_i$

Fitted model $\hat{P}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Prediction conditional on X^* $\hat{P}^* = \hat{\beta}_1 + \hat{\beta}_2 X^*$

Actual value of P^* $P^* = \beta_1 + \beta_2 X^* + u^*$

$$PE = P^* - \hat{P}^* = \left(\beta_1 + \beta_2 X^* + u^*\right) - \left(\hat{\beta}_1 + \hat{\beta}_2 X^*\right)$$

$$\begin{aligned} E(PE) &= E\left(\beta_1 + \beta_2 X^* + u^*\right) - E\left(\hat{\beta}_1 + \hat{\beta}_2 X^*\right) \\ &= \beta_1 + \beta_2 X^* + E\left(u^*\right) - E\left(\hat{\beta}_1\right) - X^* E\left(\hat{\beta}_2\right) \\ &= \beta_1 + \beta_2 X^* - \beta_1 - X^* \beta_2 \end{aligned}$$

$E(u^*) = 0$ потому что u^* выявлено случайно из распределения для u, который мы приняли как нулевую совокупность. В соответствии с обычными предположениями МНК $\hat{\beta}_1$ есть несмещенная оценка β_1 и $\hat{\beta}_2$ является несмещенной оценкой $\hat{\beta}_2$

PREDICTION

True model	$P_i = \beta_1 + \beta_2 X_i + u_i$
Fitted model	$\hat{P}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
Prediction conditional on X^*	$\hat{P}^* = \hat{\beta}_1 + \hat{\beta}_2 X^*$
Actual value of P^*	$P^* = \beta_1 + \beta_2 X^* + u^*$

$$PE = P^* - \hat{P}^* = \left(\beta_1 + \beta_2 X^* + u^*\right) - \left(\hat{\beta}_1 + \hat{\beta}_2 X^*\right)$$

$$\begin{aligned} E(PE) &= E\left(\beta_1 + \beta_2 X^* + u^*\right) - E\left(\hat{\beta}_1 + \hat{\beta}_2 X^*\right) \\ &= \beta_1 + \beta_2 X^* + E\left(u^*\right) - E\left(\hat{\beta}_1\right) - X^* E\left(\hat{\beta}_2\right) \\ &= \beta_1 + \beta_2 X^* - \beta_1 - X^* \beta_2 = 0 \end{aligned}$$

Следовательно, ожидание ошибки прогнозирования равен нулю. Результат легко обобщается в случае, где множественные характеристики и новые товары включают в себя новую комбинацию

Variance of prediction error

$$\sigma_{PE}^2 = \left\{1 + \frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\}\sigma_u^2$$

Показана дисперсия совокупности ошибки прогнозирования. Неудивительно, что это подразумевает, далее - значение X * от среднего значения выборки, что дисперсия совокупности ошибки прогнозирования будет больше.

Variance of prediction error

$$\sigma_{PE}^{2} = \left\{ 1 + \frac{1}{n} + \frac{\left(X^{*} - \overline{X}\right)^{2}}{\sum_{i=1}^{n}\left(X_{i} - \overline{X}\right)^{2}} \right\} \sigma_{u}^{2}$$

Также подразумевается, что снова неудивительно, что тем больше образец, тем меньше будет дисперсия совокупности ошибки прогнозирования с нижним лимитом σ_u^2

Variance of prediction error

$$\sigma_{PE}^2 = \left\{ 1 + \frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \right\} \sigma_u^2$$

При условии, что допущения модели регрессии действительны, $\hat{\beta}_1$ и $\hat{\beta}_2$ будут стремиться к их реальным значениям, так как образец становится больше, поэтому единственным источником ошибки в прогнозировании будет *u, и по определению оно имеет дисперсию совокупности σ_u^2.**

PREDICTION

Variance of prediction error

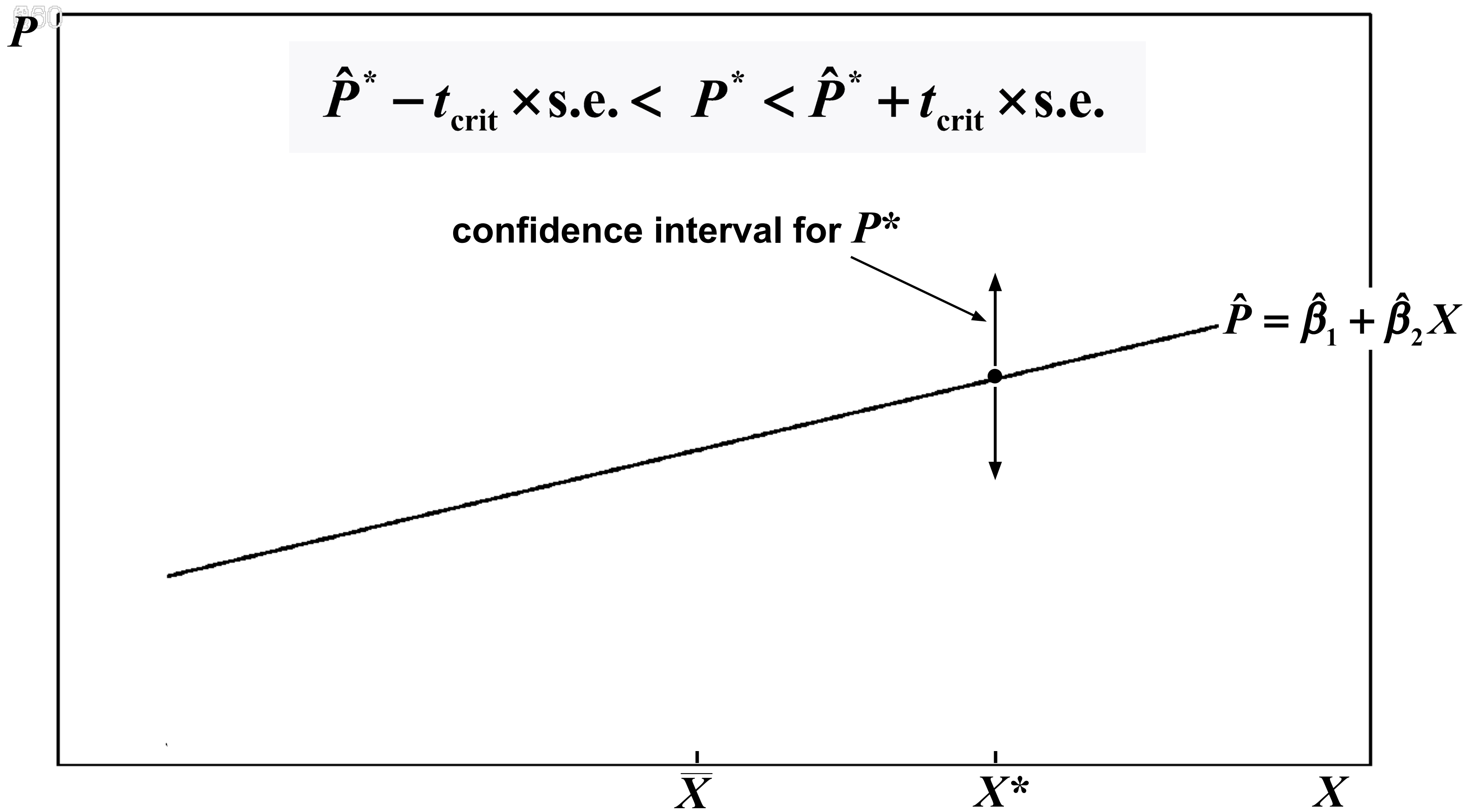
$$\sigma_{PE}^2 = \left\{1 + \frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\}\sigma_u^2$$

Standard error

$$\text{s.e.}\left(PE\right) = \sqrt{\left\{1 + \frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right\}\hat{\sigma}_u^2}$$

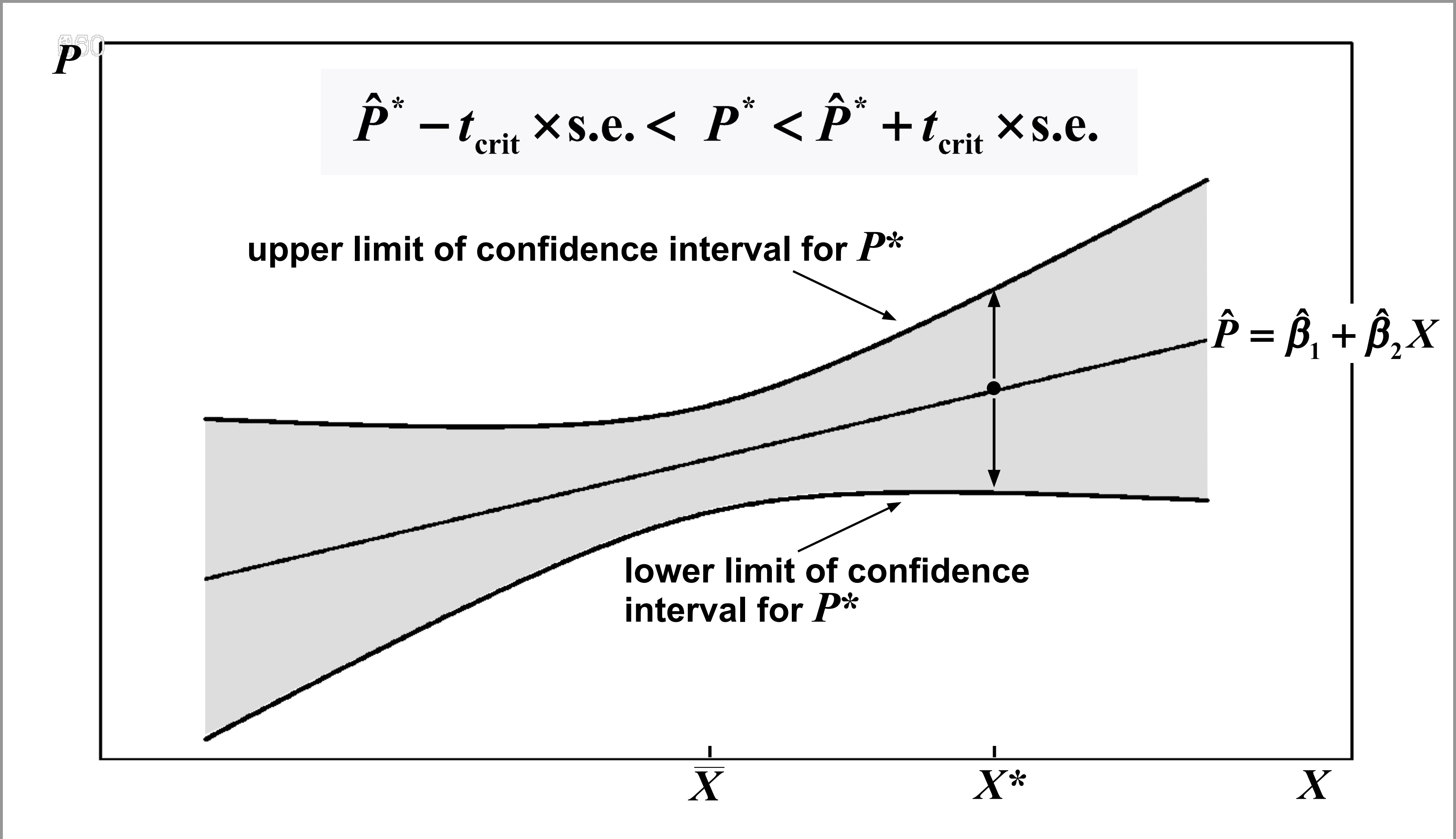
Стандартная ошибка прогнозирования вычисляется используя квадратный корень выражения для дисперсии совокупности, заменяя дисперсию u с оценкой, полученной при подгонке модели в период выборки.

$$\hat{P}^* - t_{\text{crit}} \times \text{s.e.} < P^* < \hat{P}^* + t_{\text{crit}} \times \text{s.e.}$$



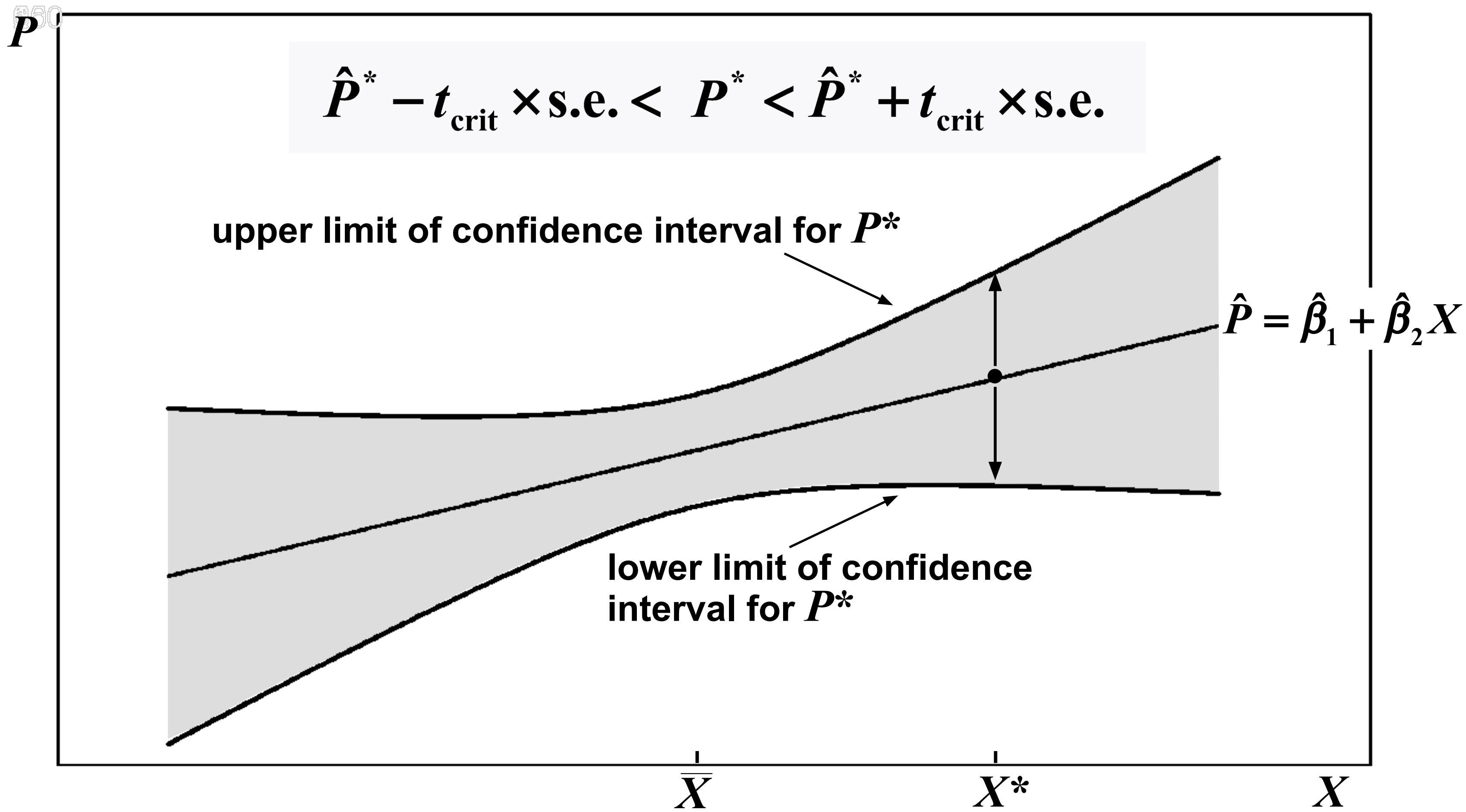
Следовательно, мы можем построить доверительный интервал для ожидания. t_{crit} критический уровень t, учитывая выбранный уровень значимости и количество степеней свободы, и s.e. является стандартной ошибкой прогнозирования.

PREDICTION



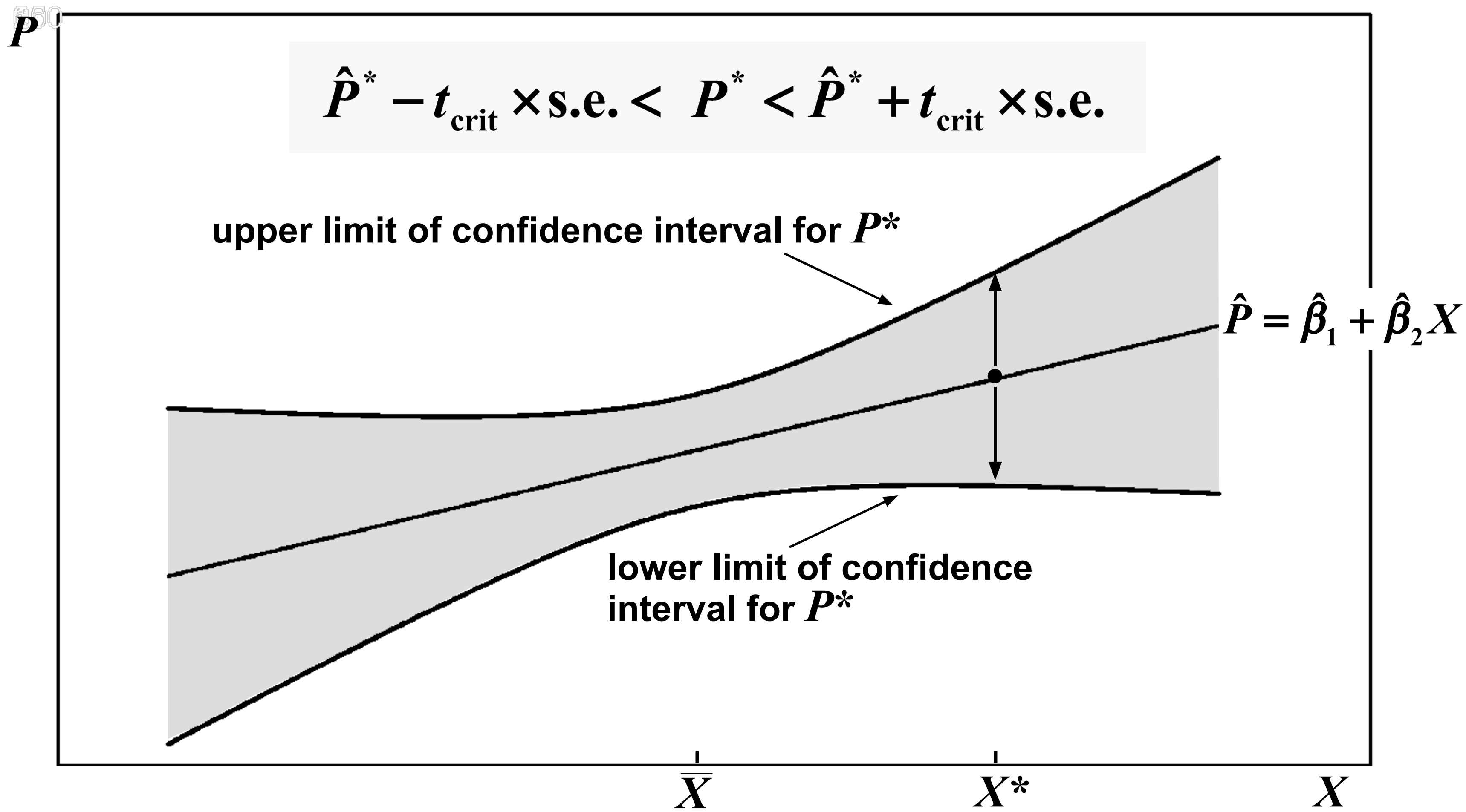
Доверительный интервал получен как функция Как мы отметили из математического выражения, он становится шире, чем больше расстояние от X * до среднего образца.

PREDICTION



С несколькими объясняющими переменными выражение для ожидание дсиперсии становится сложным

PREDICTION



Следует отметить, что мультиколлинеарность не может оказывать отрицательного влияния на точность прогнозирования, даже если оценки коэффициентов имеют большие отклонения.

PREDICTION

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

Suppose X_2 and X_3 are positively correlated, $\beta_2 > 0$, $\beta_3 > 0$.

Then $\operatorname{cov}\left(\hat{\beta}_2, \hat{\beta}_3\right) < 0$.

If $\hat{\beta}_2$ is overestimated, $\hat{\beta}_3$ is likely to be underestimated.

So $\left(\hat{\beta}_2 X_2^* + \hat{\beta}_3 X_3^*\right)$ may be a good estimator of $\left(\beta_2 X_2^* + \beta_3 X_3^*\right)$.

Similarly, for other combinations.

Для простоты предположим, что есть две объясняющие переменные, что оба имеют положительные истинные коэффициенты и как показано, они положительно коррелированы; мы прогнозируем значение Y *, учитывая значения X2 * и X3 *.

PREDICTION

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

Suppose X_2 and X_3 are positively correlated, $\beta_2 > 0$, $\beta_3 > 0$.

Then $\operatorname{cov}\left(\hat{\beta}_2, \hat{\beta}_3\right) < 0$.

If $\hat{\beta}_2$ is overestimated, $\hat{\beta}_3$ is likely to be underestimated.

So $\left(\hat{\beta}_2 X_2^* + \hat{\beta}_3 X_3^*\right)$ may be a good estimator of $\left(\beta_2 X_2^* + \beta_3 X_3^*\right)$.

Similarly, for other combinations.

Тогда, если эффект от X_2 переоценен, поэтому $\hat{\beta}_2 > \beta_2$, эффект от X_3 вероятно будет недооценена с $\hat{\beta}_3 < \beta_3$. Как следствие, эффекты ошибки могут в какой-то мере отмениться, в результате линейная комбинация может быть приближена к $(\beta_2 X_2^* + \beta_3 X_3^*)$

Simulation

$$Y = 10 + 2X_2 + 3X_3 + u$$

$$X_2 = \{1, 2, 3, 4, ..., 17, 18, 19, 20\}$$
$$X_3 = \{2, 2, 4, 4, ..., 18, 18, 20, 20\}$$

$$r_{X_2, X_3} = 0.9962$$

$$u \sim N(0,1)$$

$$Y^* = \hat{\beta}_1 + \hat{\beta}_2 X_2^* + \hat{\beta}_3 X_3^*$$
$$\cong \hat{\beta}_1 + \left(\hat{\beta}_2 + \hat{\beta}_3\right) X_2^*$$

Это будет проиллюстрировано с моделированием, с моделью и показанными данными. Мы устанавливаем модель и делаем прогноз $Y^* = \hat{\beta}_1 + \hat{\beta}_2 X_2^* + \hat{\beta}_3 X_3^*$

Simulation

$$Y = 10 + 2X_2 + 3X_3 + u$$

$$X_2 = \{1, 2, 3, 4, ..., 17, 18, 19, 20\}$$
$$X_3 = \{2, 2, 4, 4, ..., 18, 18, 20, 20\}$$

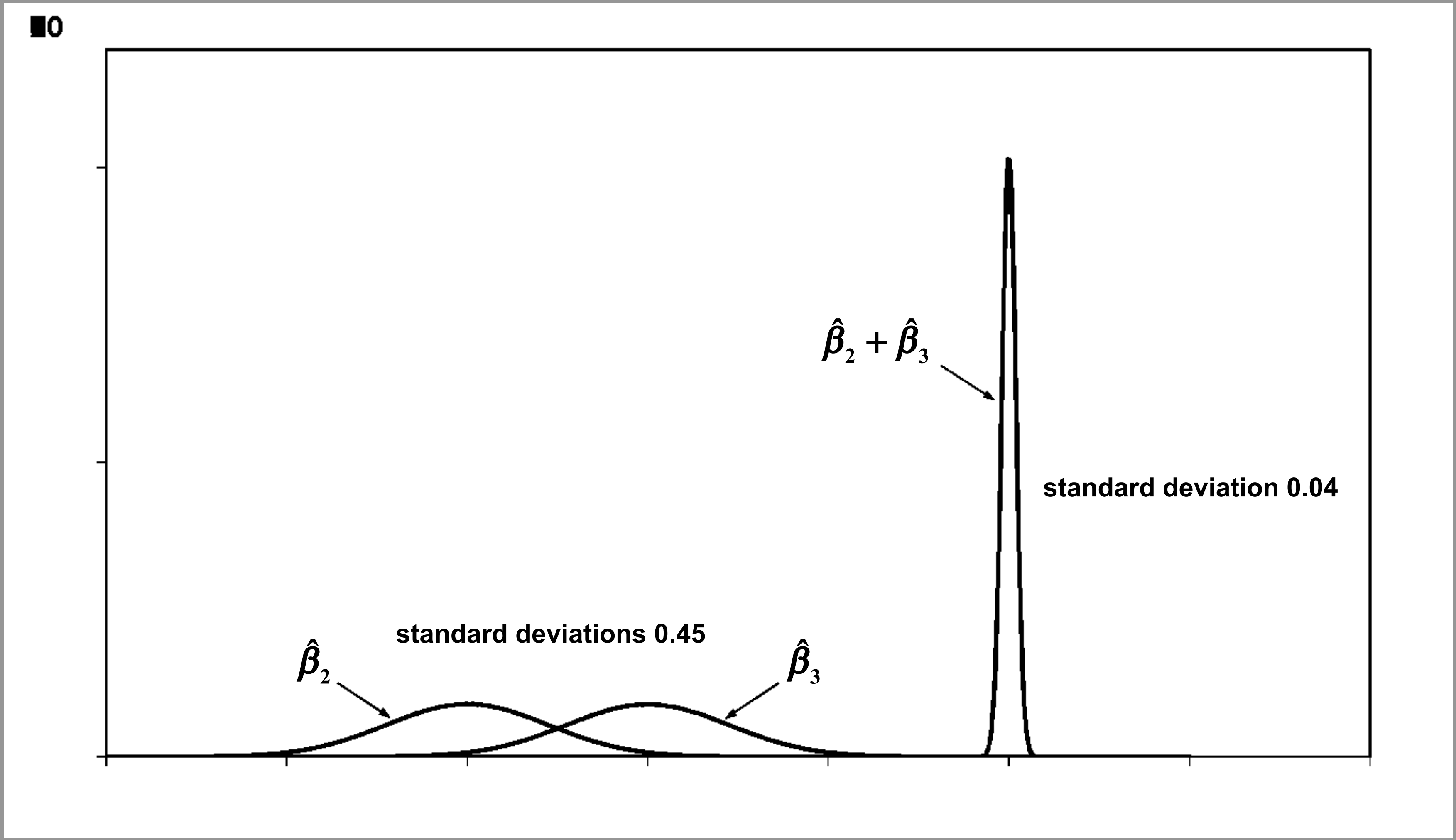
$$r_{X_2, X_3} = 0.9962$$

$$u \sim N(0,1)$$

$$Y^* = \hat{\beta}_1 + \hat{\beta}_2 X_2^* + \hat{\beta}_3 X_3^*$$
$$\cong \hat{\beta}_1 + \left(\hat{\beta}_2 + \hat{\beta}_3\right) X_2^*$$

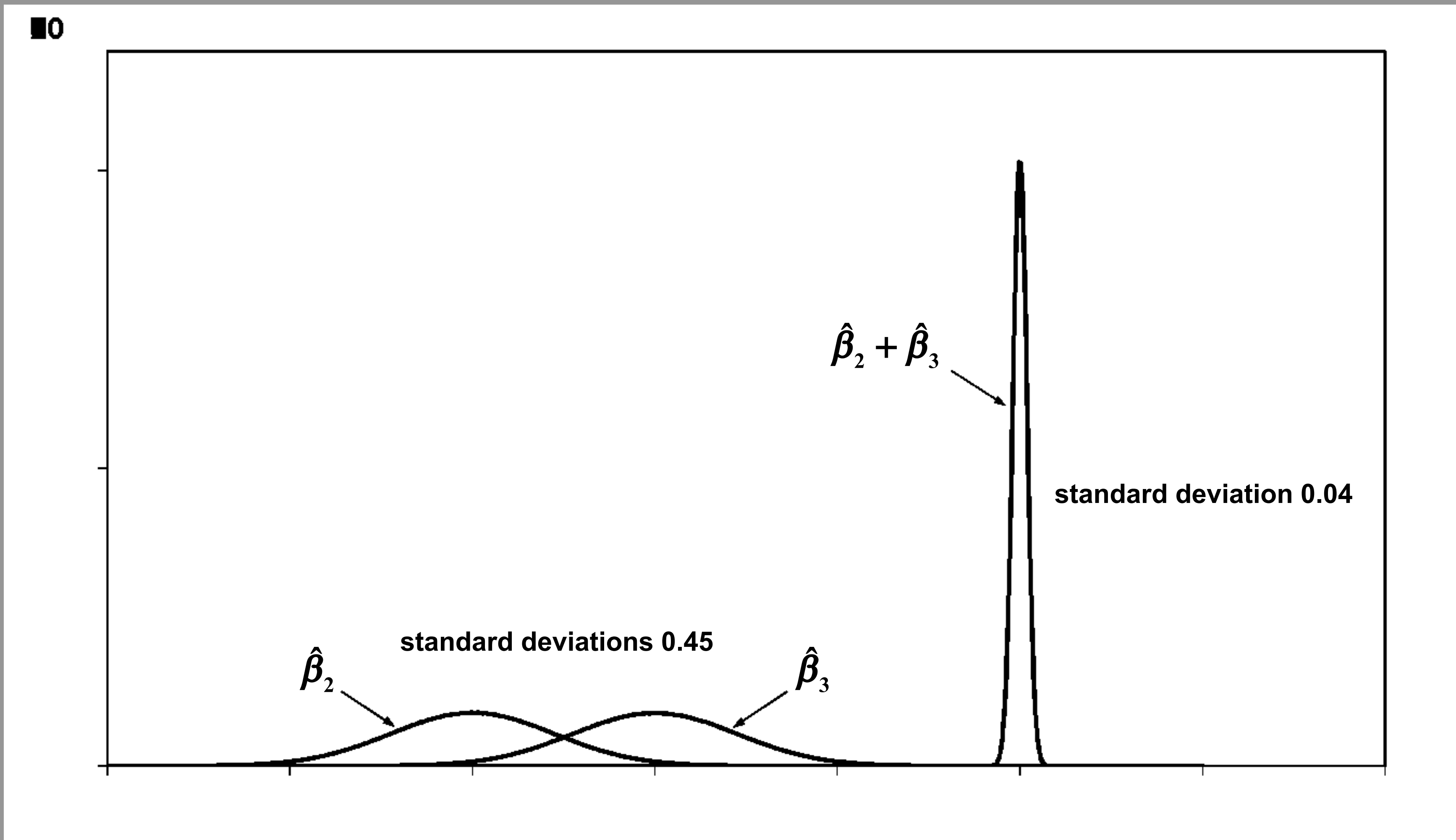
Поскольку X_2 и X_3 практически идентичны, они могут приблизиться к $Y^* = \hat{\beta}_1 + \left(\hat{\beta}_2 + \hat{\beta}_3\right) X_2^*$
Таким образом, точность прогноза зависит как близко $\left(\hat{\beta}_2 + \hat{\beta}_3\right)$ к (β_2 + β_3), то есть 5

PREDICTION



Фигура показывает распределение $\hat{\beta}_2$ и $\hat{\beta}_3$ для 10 миллионов образцов. Их распределения имеют относительно широкие отклонения вокруг их истинных моделей, как и следовало ожидать, учитывая мультиколлинеарность. Фактические стандартные отклонения их распределений составляют 0,45.

PREDICTION



Фигура также показывает сумму их распределения. Как и ожидалось, он распределяется примерно на 5, но с гораздо более низким стандартным отклонением 0,04 , несмотря на мультиколлинеарность, влияющую на точечные оценки отдельных коэффициентов