Gov 2001 Section 11, 2025

OLS Interval Estimation

1. The foundation of constructing an asymptotic CI of β is the asymptotic normality of $\hat{\beta}$ with large sample. We have

$$\hat{\beta} - \beta = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i e_i\right)$$

Both components are in the form of sample mean, so by LLN and CMT, we have

$$\hat{\beta} \xrightarrow{p} \beta + (E[\mathbf{X}_i \mathbf{X}_i'])^{-1} E[\mathbf{X}_i e_i] = \beta$$

And we can use CLT to find the asymptotic distribution of the second component:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i e_i - E[\mathbf{X}_i e_i]\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{X}_i e_i \xrightarrow{d} \mathcal{N}(0, E[e_i^2 \mathbf{X}_i \mathbf{X}_i'])$$

Since $\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' \xrightarrow{p} E[\mathbf{X}_i \mathbf{X}_i']$ and

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{X}_i e_i\right)$$

By Slutsky's theorem, we have

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_\beta), \quad \mathbf{V}_\beta = (\mathbb{E}[\mathbf{X}_i \mathbf{X}_i'])^{-1} \mathbb{E}[e_i^2 \mathbf{X}_i \mathbf{X}_i'] (\mathbb{E}[\mathbf{X}_i \mathbf{X}_i'])^{-1}$$

2. The variance of $\hat{\beta}$'s asymptotic distribution contains expectations, which are not directly observable in real life, so we need to estimate them. The most straightforward method is again plug-in estimators, replacing population means with sample means:

$$\hat{\mathbf{V}}_\beta = \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{e_i}^2 \mathbf{X}_i \mathbf{X}_i'\right) \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'\right), \quad \hat{e_i} = Y_i - \mathbf{X}_i' \hat{\beta}$$

This is called the robust variance estimator, as it does not assume $\text{Var}(e_i^2 | \mathbf{X}_i) = \text{Var}(e_i^2) = \sigma^2$. If we assume homoskedasticity, we have by Law of iterated expectations,

$$\mathbb{E}[e_i^2 \mathbf{X}_i \mathbf{X}_i'] = \mathbb{E}[\mathbb{E}[e_i^2 | \mathbf{X}_i] \mathbf{X}_i \mathbf{X}_i'] = \mathbb{E}[e_i^2] \mathbb{E}[\mathbf{X}_i \mathbf{X}_i'] = \sigma^2 \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'\right)^{-1}.$$

We just need to estimate σ^2, replacing it by $\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} \hat{e}_i^2$. Plugging in to $\mathbf{V}_\beta$, the homoskedasticity (standard) variance estimator is just

$$\hat{\mathbf{V}}_\beta^{lm} = \hat{\sigma}^2 \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'\right)^{-1}$$

The robust and standard SEs are just $\sqrt{\hat{\mathbf{V}}_\beta / n}$ and $\sqrt{\hat{\mathbf{V}}_\beta^{lm} / n}$ respectively.

3. For each $\hat{\beta}_j$ in $\hat{\beta}$, we test $H_0 : \hat{\beta}_j = 0$ versus $H_1 : \hat{\beta}_j \neq 0$. Under the null, we have

$$\frac{\hat{\beta}_j - 0}{\hat{\text{se}}(\hat{\beta}_j)} \xrightarrow{d} \mathcal{N}(0,1), \quad \hat{\text{se}}(\hat{\beta}_j) = \sqrt{\frac{[\hat{\mathbf{V}}_\beta]_{jj}}{n}}$$

using robust SE (similar for standard SE). Therefore, we can invert the test and easily get 95% CI:

$$\left[\hat{\beta}_j - 1.96\hat{\text{se}}(\hat{\beta}_j), \hat{\beta}_j + 1.96\hat{\text{se}}(\hat{\beta}_j)\right]$$

Practice Questions

1. In recent years it has become common in statistics to want to perform many simultaneous hypothesis tests. Let $p_1 \dots p_m$ be indepedent p-values, corresponding to m hypothesis tests. Each of the m hypothesis tests has a simple null. Suppose that m_0 of the m null hypotheses are true. We decide in advance to conduct these tests at level α (i.e., we reject the null for tests where the p-value is less than α). The *familywise error rate* is the probability of making at least one Type I error.

   (a) Find the familywise error rate (FWER). What happens to the familywise error rate as m_0 gets large? (**Hint:** Let V be the *number* of Type I error rates. Then FWER is $P(V > 0)$. You can simplify it further by using α and m_0.)

   (b) One of the common procedure to deal with multiple hypothesis testings is called the *Bonferroni procedure*. Intuitively, this procedure allows us to deal with the increased likelihood of type I errors. This procedure is described as follows: instead of rejecting the null hypotheses with $p_i < \alpha$, we reject the null hypotheses with $p_i < \frac{\alpha}{m}$ (i.e., correcting the cutoff by the number of hypotheses). Show that under this procedure, the familywise error rate is at most α. (**Hint:** You might find Markov's Inequality helpful: $\text{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$. You want to bound the FWER using Markov's Inequality by setting the value of a appropriately.)

   (c) In (b), why not instead reject the null hypotheses with $p_i < \frac{\alpha}{m_0}$, considering that $m_0 \leq m$ (and often in practice m_0 is much smaller than m), which would seem to result in rejecting more false nulls while still keeping the familywise error rate at most α?

   (d) Another procedure is to reject all null hypotheses with $p_i < 1 - (1-\alpha)^{1/m}$ (This is known as the Sidak Procedure). Show that under this procedure, the familywise error rate is again at most α.

2. Often our data is collected with error, which we refer to as measurement error. For instance, for a dependent variable Y you're trying to measure in a survey, respondents may randomly mis-click, or they may systematically lie about having a socially undesirable trait. In this question, we will explore the impact of measurement error in regression analysis in the most favourable case where the measurement error is independent of the true values. Consider the linear projection:

$$L[Y \mid 1, X] = \beta_0 + \beta_1 X$$

with the projection error denoted as $e = Y - L[Y \mid 1, X]$, and $\mathbb{V}[X] = \sigma_X^2$. Unfortunately, we do not observe Y or X but instead noisy proxies for them $\{\tilde{Y}, \tilde{X}\}$, where

$$\tilde{Y} = Y + v, \quad \tilde{X} = X + w$$

Where v is one realization from $V \sim \mathcal{N}(0, \sigma_v^2)$ and w is one realization from $W \sim \mathcal{N}(0, \sigma_w^2)$, where W and V are independent of X and Y. This implies that $\text{Cov}(v, X) = \text{Cov}(v, e) = \text{Cov}(v, w) = \text{Cov}(w, X) = \text{Cov}(w, e) = \text{Cov}(w, v) = 0$. This is commonly referred to as classical measurement error.

(a) Consider the linear projection of these observable variables, $L[\tilde{Y} \mid 1, \tilde{X}] = \alpha_0 + \alpha_1 \tilde{X}$. Find α_1 in terms of $\{\beta_1, \sigma_w^2, \sigma_v^2, \sigma_X^2\}$. Hint: first derive an expression of the coefficients in terms of the $\tilde{X}$ and $\tilde{Y}$.

(b) From your expression in part (a), briefly explain (1-2 sentences) the effect of this type of measurement error in X on the sign and magnitude of the coefficient α_1 compared to β_1. Hint: what parameter controls the amount of measurement error in X?

(c) From your expression in part (a), briefly explain (1-2 sentences) the effect of this type of measurement error in Y on the sign and magnitude of the coefficient α_1 compared to β_1. Hint: what parameter controls the amount of measurement error in Y?

3. Decide whether each of the following statements are true or false and explain your reasoning briefly.

   (a) If $Y = X\beta + e$, $X \in \mathbb{R}$, and $E[e|X] = 0$, then $E[e] = 0$.

   (b) If $Y = X\beta + e$, $X \in \mathbb{R}$, and $E[e|X] = 0$, then $E[X^3 e] = 0$.

   (c) If $Y = X\beta + e$, $X \in \mathbb{R}$, and $E[Xe] = 0$, then $E[X^2 e] = 0$.

   (d) If $Y = X\beta + e$, $X \in \mathbb{R}$, and $E[e|X] = 0$, then e and X are independent.

4. In most linear regression models, the dependent variable Y is expressed as a function of independent variables $X_1, X_2, ..., X_k$ (or to use the vector notation, just X as a vector in $\mathbb{R}^k$). That is,

$$Y = X\beta + e$$

   where β is a $k \times 1$ coefficient vector and e is the error.

   (a) Explain briefly what it means for $g(X)$, a function of X, to be the best predictor of Y.

   (b) Show that for $g(X)$ to be the best predictor, $g(X)$ must be equal to the conditional expectation function $E[Y|X]$. (**Hint:** You can assume that, for the CEF error $e = Y - E[Y|X]$, we have $E\big|e\{E[Y|X] - g(X)\}\big| < \infty$).

   Now, for unifying definition, we also need to consider an *intercept-only* model, where there is no X and α is simply a constant:

$$Y = \alpha + e$$

   (c) Find $\underset{\alpha}{\text{argmin}} E[(Y - \alpha)^2]$.

5. Consider the following multivariate regression:

$$Y = X\beta + Z\gamma + \epsilon$$

(a) Show that for any $\{X, Z\}$, we can decompose Z into $P_X Z + M_X Z$, where P_X and M_X are the projection matrix and annihilator matrix of X respectively. (**Hint**: Apply the definitions of the projection and annihilator matrices.) Also show that P_X and M_X are orthogonal. (**Hint**: Show that $P_X^T M_X = 0$.)

(b) Show that if $X \perp Z$, then the coefficients we get from regressing Y on X and Y on Z will be the same coefficients from the joint regression above. (**Hint**: One way to work on this problem is to notice that $Y = X\beta + Z\gamma + \epsilon$, and thus $Cov(Y, X) = Cov(X\beta + Z\gamma + \epsilon, X)$.)

(c) Suppose $\hat{\beta}$ and $\hat{\gamma}$ are the OLS estimators for β and γ for the regression above. Find a $\hat{\beta}'$ such that:

$$\hat{Y} = X\hat{\beta}' + (M_X Z)\hat{\gamma}$$

Write $\hat{\beta}'$ in terms of $\hat{\beta}$ and $\hat{\gamma}$, and provide a substantive interpretation of $\hat{\beta}'$ in plain English (**Hint**: X and Z are not necessarily orthogonal anymore. Use results from part (a).).

(d) Lastly, show that the following regression

$$M_X \hat{Y} = (M_X Z)\hat{\gamma}$$

will return the same OLS estimator $\hat{\gamma}$ as in the multivariate regression $Y = X\beta + Z\gamma + \epsilon$, explain this result in plain English.

6. The standard output from OLS will give the standard errors for the estimated coefficients, but often we want to obtain measures of uncertainty for the predicted value of Y_i given some value of X_i (that is, the conditional expectation function). Using the example from lecture, we might be interested in the average wait times to vote for individuals making \$25,000, \$50,000, or \$100,000 in annual income, along with measures of uncertainty around those estimates. In this problem we will look at how to calculate interval estimates for these predicted values. Assume the following *true* population model for $Y_i|X_i$:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

where the X_i are random variables and u_i are i.i.d. random variables with $E[u_i \mid X_i] = 0$ and $Var(u_i \mid X_i) = \sigma^2$. Suppose we observe a random sample of n paired observations $\{Y_i, X_i\}$. Assume the Gauss-Markov assumptions hold (i.e., the OLS estimator is unbiased) and that we have a large sample. Our goal is to estimate the predicted value at some value $X_i = x$:

$$\mu(x) = E[Y_i \mid X_i = x] = \beta_0 + \beta_1 x.$$

(a) Let $\hat{\beta}_0$ and $\hat{\beta}_1$ be OLS estimators of the regression of Y on X. Use what you know about the unbiasedness of OLS estimates to show that $\widehat{\mu}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ is an unbiased estimator of the population quantity $\mu(x) = E[Y_i \mid X_i = x]$.

(b) Find the conditional variance of $\hat{\beta}_0$, $Var(\hat{\beta}_0 \mid X_1, \ldots, X_n)$, using the following two facts. You answer should be in terms of σ^2 and functions of X_i.

$$\text{Cov}(\overline{Y}, \hat{\beta}_1 \mid X_1, \ldots, X_n) = 0 \quad \text{and} \quad Var(\hat{\beta}_1 \mid X_1, \ldots, X_n) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}.$$

(c) Find the covariance of the OLS estimates given our X values, $\text{Cov}(\hat{\beta}_0, \hat{\beta}_1 | X_1, \ldots, X_n)$, again in terms of σ^2 and functions of the X_i. (**Hint**: It's not zero.)

(d) Using what you found in (b) and (c), find the standard error of $\widehat{\mu}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$.

(e) Assume that we don't know σ^2 and instead construct our estimate of the standard error by plugging in for σ^2 our unbiased estimate s^2 using the residuals. Give the formula for a large-sample 95% confidence interval estimator for $\mu(x) = E[Y \mid X = x]$ using what you found above and substituting s^2 for σ^2. How do we interpret this confidence interval?

Practice Question Solutions

1. (a) Let V be the number of Type I errors, i.e., the number of true null hypotheses incorrectly rejected. Then the FWER is:

$$\text{FWER} = \mathbb{P}(V > 0)$$

Under the assumption that the m_0 true nulls are independent and each test is conducted at level α, the probability that a single true null hypothesis is *not* rejected is $1 - \alpha$. Hence, the probability that *none* of the m_0 true nulls are rejected is:

$$\mathbb{P}(V = 0) = (1 - \alpha)^{m_0}$$

Therefore, the familywise error rate is:

$$\text{FWER} = 1 - (1 - \alpha)^{m_0}$$

As m_0 increases, $(1 - \alpha)^{m_0} \to 0$, so:

$$\text{FWER} \to 1$$

That is, the FWER increases with the number of true null hypotheses, and we almost surely would make one or more false discoveries.

(b) Let V again denote the number of Type I errors. Since each true null has probability at most $\frac{\alpha}{m}$ of being rejected, the expected number of false rejections is:

$$\mathbb{E}[V] \leq m_0 \cdot \frac{\alpha}{m} \leq \alpha$$

since $m_0 \leq m$. By Markov's Inequality:

$$\mathbb{P}(V \geq 1) \leq \mathbb{E}[V] \leq \alpha$$

Thus, under the Bonferroni procedure:

$$\text{FWER} = \mathbb{P}(V > 0) \leq \alpha$$

(c) While using $\frac{\alpha}{m_0}$ as the cutoff would indeed increase power (since $m_0 \leq m$), the issue is that m_0 is unknown in practice. If we underestimate m_0, we risk violating the FWER control (i.e., making it exceed α). Thus, we use $\frac{\alpha}{m}$ to ensure valid FWER control without knowing which nulls are true. However, this also makes the Bonferroni procedure extremely conservative.

(d) Under the assumption of independence, the probability that all m_0 true null hypotheses are not rejected is:

$$\mathbb{P}(V = 0) = (1 - \alpha)^{m_0/m}$$

So the FWER becomes:

$$\text{FWER} = 1 - (1 - \alpha)^{m_0/m} \leq \alpha$$

since $(1 - \alpha)^{m_0/m} \geq 1 - \alpha$ when $m_0 \leq m$. Therefore, the Sidak procedure also controls the familywise error rate at level α under independence.

2. (a) Substituting the original Y and X by the noisy observation, we have

$$\tilde{Y} = \beta_0 + \beta_1 \tilde{X} - \beta_1 w + e + v$$

Now, we know that

$$\alpha_1 = \frac{\text{Cov}[\tilde{X}, \tilde{Y}]}{V[\tilde{X}]} = \frac{\text{Cov}[X + w, \beta_0 + \beta_1 X + e + v]}{V[X + w]} = \frac{\beta_1 \text{Cov}[X, X]}{\sigma_X^2 + \sigma_w^2} = \frac{\beta_1 \sigma_X^2}{\sigma_X^2 + \sigma_w^2}$$

(b) The measurement error w biases the estimate of the coefficient. When $\sigma_w^2 > 0$, we have $\alpha_1 < \beta_1$, and β_1 is biased towards zero by $\frac{\beta_1 \sigma_X^2}{\sigma_X^2 + \sigma_w^2}$. This is because we introduced additional variance to the predictor.

(c) The measurement error v does not bias the estimate of the coefficient. It is uncorrelated with other variables and just causes less precision when estimating the mean effect.

3. (a) **True.** If $\mathbb{E}[e \mid X] = 0$, then taking the expectation over X gives $\mathbb{E}[e] = \mathbb{E}[\mathbb{E}[e \mid X]] = 0$ by the Law of Iterated Expectations.

(b) **True.** Again using the Law of Iterated Expectations, and remember that $\mathbb{E}[g(X)|X] = g(X)$ for any function g:

$$\mathbb{E}[X^3 e] = \mathbb{E}[\mathbb{E}[X^3 e \mid X]] = \mathbb{E}[X^3 \cdot \mathbb{E}[e \mid X]] = \mathbb{E}[X^3 \cdot 0] = 0.$$

(c) **False.** $\mathbb{E}[Xe] = 0$ does not imply $\mathbb{E}[X^2 e] = 0$. The expectation $\mathbb{E}[X^2 e]$ involves a different function of X and may not be zero unless $\mathbb{E}[e \mid X] = 0$. Counterexample:

Let $X \sim \mathcal{N}(0,1)$, and define $e = X^2 - 1$. Then:

$$\mathbb{E}[Xe] = \mathbb{E}[X(X^2-1)] = \mathbb{E}[X^3] - \mathbb{E}[X] = 0 - 0 = 0.$$

So $\mathbb{E}[Xe] = 0$. But

$$\mathbb{E}[X^2 e] = \mathbb{E}[X^2(X^2-1)] = \mathbb{E}[X^4] - \mathbb{E}[X^2] = 3 - 1 = 2 \neq 0.$$

(d) **False.** $\mathbb{E}[e \mid X] = 0$ implies mean independence, not full independence. e and X can still be dependent in higher moments or in distribution. Stochastic independence implies mean independence, but the converse is not true; mean independence implies uncorrelatedness, while the converse is not true.

4. (a) A function $g(X)$ is the best predictor of Y if it minimizes the expected squared prediction error, i.e.,

$$g(X) = \arg\min_{f(X)} \mathbb{E}[(Y - f(X))^2].$$

(b) Let $e = Y - \mathbb{E}[Y|X]$ be the CEF error. For any function $g(X)$, we have:

$$\begin{aligned}\mathbb{E}[(Y-g(X))^2] &= \mathbb{E}[(\mathbb{E}[Y|X] + e - g(X))^2] \\ &= \mathbb{E}[(\mathbb{E}[Y|X] - g(X))^2] + 2\mathbb{E}[e(\mathbb{E}[Y|X] - g(X))] + \mathbb{E}[e^2]\end{aligned}$$

Since $\mathbb{E}[e|X] = 0$, the cross term vanishes by the Law of Iterated Expectations (remember that CEF given X is a function of X):

$$\mathbb{E}[e(\mathbb{E}[Y|X] - g(X)) \mid X] = \mathbb{E}[\mathbb{E}[e|X](\mathbb{E}[Y|X] - g(X))] = 0.$$

Thus,

$$\mathbb{E}[(Y-g(X))^2] = \mathbb{E}[(\mathbb{E}[Y \mid X] - g(X))^2] + \mathbb{E}[e^2],$$

which is minimized when $g(X) = \mathbb{E}[Y \mid X]$. Therefore, the CEF is the best predictor.

(c) We want to minimize $\mathbb{E}[(Y-\alpha)^2]$ over α. Taking the derivative:

$$\frac{d}{d\alpha}\mathbb{E}[(Y-\alpha)^2] = \mathbb{E}[-2(Y-\alpha)] = -2(\mathbb{E}[Y] - \alpha).$$

Setting to zero gives $\alpha = \mathbb{E}[Y]$. So the best constant predictor is the mean:

$$\arg\min_{\alpha} \mathbb{E}[(Y-\alpha)^2] = \mathbb{E}[Y].$$

5. (a) By definition, the projection matrix $P_X = X(X'X)^{-1}X'$, and the annihilator matrix is $M_X = I - P_X$. So for any Z,

$$Z = IZ = (P_X + I - P_X)Z = P_X Z + M_X Z,$$

which decomposes Z into its projection onto the column space of X and its residual.

To show orthogonality, since projection matrices are symmetric and idempotent,

$$P_X^T M_X = P_X M_X = P_X(I - P_X) = P_X - P_X = 0.$$

Thus, P_X and M_X are orthogonal.

(b) If $X \perp Z$, then $\text{Cov}(X, Z) = 0$. From the regression:

$$Y = X\beta + Z\gamma + \epsilon,$$

we have:

$$\text{Cov}(Y, X) = \text{Cov}(X\beta + Z\gamma + \epsilon, X) = \beta\text{Cov}(X, X) + \gamma\text{Cov}(Z, X) + \text{Cov}(\epsilon, X).$$

Since $X \perp Z$ and $E[\epsilon] = E[X\epsilon] = 0$, the last two terms are zero, and we get the same result as in the simple regression of Y on X alone:

$$\text{Cov}(Y, X) = \text{Cov}(\beta X + \varepsilon, X) = \beta\text{Cov}(X, X) + \text{Cov}(\varepsilon, X) = \beta\text{Var}(X)$$

A similar argument applies to Z. Hence, the marginal regressions recover the same coefficients as in the joint regression.

(c) $\hat{Y} = X\hat{\beta} + Z\hat{\gamma}$. From the decomposition $Z = P_X Z + M_X Z$, we can write:

$$Z\hat{\gamma} = P_X Z\hat{\gamma} + M_X Z\hat{\gamma}.$$

So:

$$\hat{Y} = X\hat{\beta} + P_X Z\hat{\gamma} + M_X Z\hat{\gamma}.$$

Note that $P_X Z = X(X'X)^{-1}X'Z$, we have

$$\hat{Y} = X\left(\hat{\beta} + (X'X)^{-1}X'Z\hat{\gamma}\right) + (M_X Z)\hat{\gamma}.$$

Thus, define:

$$\hat{\beta}' = \hat{\beta} + (X'X)^{-1}X'Z\hat{\gamma}.$$

Interpretation: $\hat{\beta}'$ captures both the direct effect of X on Y, i.e., $\hat{\beta}$, and the part of Z's effect on X (the second term).

(d) Consider:

$$M_X\hat{Y} = M_X(X\hat{\beta} + Z\hat{\gamma}) = M_X Z\hat{\gamma},$$

since $M_X X = 0$ by the property of annihilator matrix. So regressing $M_X\hat{Y}$ on $M_X Z$ yields $\hat{\gamma}$. This result shows that the coefficient on Z in the multivariate regression is the same as the coefficient obtained when projecting both Y and Z orthogonally to X. That is, $\hat{\gamma}$ represents the effect of Z on Y after removing the influence of X from both.

6. (a) Given Gauss-Markov, we already know that $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased estimators, so

$$\mathbb{E}[\mu(x) \mid X_i = x] = \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 X_i \mid X_i = x] = \mathbb{E}[\hat{\beta}_0 \mid X_i = x] + \mathbb{E}[\hat{\beta}_1 X_i \mid X_i = x] = \beta_0 + \beta_1 x$$

(b) Noting that the two hints include sample averages of X and Y, let's try to write our estimated model (denoting $\hat{\varepsilon}_i$ as the residual in our estimated model, and all quantities with $\bar{*}$ as the sample averages):

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\varepsilon}_i,$$

Thus,

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} + \bar{\varepsilon}$$

Notice that $\bar{\varepsilon}$ is mechanically zero (remember the first-order condition when deriving $\hat{\beta}$ in the second approach in Section 10 is just $\sum_{i=1}^n \hat{\varepsilon}_i = 0$).

Now, let's take the conditional variance by subbing in for $\hat{\beta}_0$ (denoting $X_1, \ldots, X_n$ as X):

$$\begin{aligned}
\text{Var}(\hat{\beta}_0 \mid X) &= \text{Var}(\bar{Y} - \bar{X}\hat{\beta}_1 \mid X) \\
&= \text{Var}(\bar{Y} \mid X) + \text{Var}(\bar{X}\hat{\beta}_1 \mid X) - 2\text{Cov}(\bar{Y}, \bar{X}\hat{\beta}_1 \mid X) \\
&= \text{Var}(\bar{Y} \mid X) + \bar{X}^2 \cdot \text{Var}(\hat{\beta}_1 \mid X) - 2\bar{X} \cdot \text{Cov}(\bar{Y}, \hat{\beta}_1 \mid X) \\
&= \text{Var}(\bar{Y} \mid X) + \bar{X}^2 \cdot \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \text{Var}\left(\frac{1}{n}\sum_{i=1}^n Y_i \mid X\right) + \frac{\bar{X}^2\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \\
&= \frac{\sigma^2}{n} + \frac{\bar{X}^2\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}
\end{aligned}$$

This is tricky but note that even though $\mathbb{E}[u] = 0$, it is not true that $\sum_{i=1}^n u_i = 0$. There's a difference between the expectation of a random variable (the former) and the average of n finite draws of that variable (the latter)!

(c) We have

$$\begin{aligned}
\text{Cov}(\hat{\beta}_0, \hat{\beta}_1 \mid X) &= \text{Cov}(\bar{Y} - \bar{X}\hat{\beta}_1, \hat{\beta}_1 \mid X) \\
&= \text{Cov}(\bar{Y}, \hat{\beta}_1 \mid X) - \bar{X} \cdot \text{Cov}(\hat{\beta}_1, \hat{\beta}_1 \mid X) \\
&= -\frac{\bar{X}\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}
\end{aligned}$$

(d) Once we start expanding out this variance expression, we quickly find we have all the

ingredients we need:

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}(x) \mid X) &= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 x \mid X) \\
&= \mathrm{Var}(\hat{\beta}_0 \mid X) + x^2 \cdot \mathrm{Var}(\hat{\beta}_1 \mid X) + 2x \cdot \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1 \mid X) \\
&= \frac{\sigma^2}{n} + \frac{\bar{X}^2\sigma^2}{\sum(X_i - \bar{X})^2} + \frac{x^2\sigma^2}{\sum(X_i - \bar{X})^2} - \frac{2x\bar{X}\sigma^2}{\sum(X_i - \bar{X})^2} \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2(\bar{X}^2 + x^2 - 2x\bar{X})}{\sum(X_i - \bar{X})^2} \\
&= \frac{\sigma^2}{n} + \frac{\sigma^2(x - \bar{X})^2}{\sum(X_i - \bar{X})^2}
\end{aligned}$$

So the standard error is:

$$\mathrm{SE}(\hat{\mu}(x) \mid X) = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2(x - \bar{X})^2}{\sum(X_i - \bar{X})^2}}$$

(e) Since we don't know σ^2 in data analysis, we estimate it with the residual variance:

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

So the estimated standard error is:

$$\widehat{\mathrm{SE}}(\hat{\mu}(x)) = \sqrt{\frac{\hat{\sigma}^2}{n} + \frac{\hat{\sigma}^2(x - X)^2}{\sum(X_i - \bar{X})^2}}$$

The asymptotic distribution is:

$$\sqrt{n} \cdot \frac{(\hat{\mu}(x) - \mu(x))}{\widehat{\mathrm{SE}}(\hat{\mu}(x))} \xrightarrow{d} \mathcal{N}(0, 1)$$

So the large-sample 95% confidence interval is:

$$\hat{\mu}(x) \pm 1.96 \cdot \widehat{\mathrm{SE}}(\hat{\mu}(x))$$

Interpretation: In repeated large samples, 95% of such confidence intervals will contain the true conditional expectation of Y at a given x, assuming we estimate OLS from an i.i.d. sample $\{X_1, \ldots, X_n\}$.