HE2003 Econometrics I

Tutorial 4 (Week 5) – Problems & Solutions

Academic Year 2026/2027, Semester 1

Quantitative Research Society @NTU

October 12, 2026

Overview

- **Lecture coverage:** Lecture 4 — *Linear Regression with One Regressor 3*: further selection-bias examples (hospitals, wages, program evaluation, surveys, ratings); Least Squares Assumption 2 (i.i.d. sampling) and why simple random sampling matters; the measure of fit R^2 and its ESS/TSS decomposition; the Capital Asset Pricing Model (CAPM) and the "beta" of a stock as a running OLS application.
- **Questions:** 4 questions from Tutorial Question Set No. 4.
- **Notation:** as in lectures — u_i population error, $\hat{u}_i$ OLS residual; ESS explained sum of squares, SSR residual sum of squares, TSS = ESS + SSR; "Least Squares Assumption 1/2/3"; critical value 1.96 for a two-sided 5% test.
- **Question themes:** Q1 selection-bias examples and the measure of fit R^2; Q2 predictions from a weight–height regression; Q3 consequences of $\hat{\beta}_1 = 0$ for $\hat{Y}_i$ and R^2; Q4 return and risk of CAPM assets with equal β and different α.

For academic reference only. This document is provided for educational purposes and does not constitute an official question paper, answer key, or any formally authorised academic material. Its content is not guaranteed to be complete or accurate.

Question 1 (Selection bias examples; the measure of fit R^2)

Problem

(a) We discussed the selection bias/survivorship bias in class. Can you give some other examples of this bias?

(b) Explain the measure of fit R^2. Illustrate the meaning of explained sum of squares (ESS) and total sum of squares (TSS) by using a diagram.

(Tutorial Question Set No. 4, Q1.)

Solution

Part (a), Method 1: Least Squares Assumption 1 diagnosis

Recall Least Squares Assumption 1: $E[u_i \mid X_i] = 0$. By the law of iterated expectations (LIE) it implies that the error has mean zero and is uncorrelated with the regressor:

$$\begin{aligned} E[u_i] &= E\big[E[u_i \mid X_i]\big] = E[0] = 0, \\ \mathrm{Cov}(X_i, u_i) &= E[X_i u_i] - E[X_i]\,E[u_i] \\ &= E\big[X_i\,E[u_i \mid X_i]\big] - E[X_i] \cdot 0 \\ &= E[X_i \cdot 0] = 0. \end{aligned}$$

Hence, by the contrapositive, any selection mechanism that makes $\mathrm{Cov}(X_i, u_i) \neq 0$ forces $E[u_i \mid X_i] \neq 0$, and Least Squares Assumption 1 fails.

The direction of the bias follows from a decomposition of $\hat{\beta}_1$. Average the model $Y_i = \beta_0 + \beta_1 X_i + u_i$ over i to get $\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$ with $\bar{u} = \frac{1}{n}\sum_{i=1}^n u_i$, and subtract:

$$Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u}).$$

Multiply by $(X_i - \bar{X})$ and sum:

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) &= \beta_1 \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (X_i - \bar{X})(u_i - \bar{u}) \\ &= \beta_1 \sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^n (X_i - \bar{X}) u_i - \bar{u} \sum_{i=1}^n (X_i - \bar{X}). \end{aligned}$$

The last sum is zero:

$$\sum_{i=1}^n (X_i - \bar{X}) = \sum_{i=1}^n X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0.$$

Divide by $\sum_{i=1}^n (X_i - \bar{X})^2$:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \beta_1 + \frac{\sum_{i=1}^n (X_i - \bar{X}) u_i}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

The second term has the sign of the sample covariance between X_i and u_i. When selection makes this covariance positive in the observed sample, $\hat{\beta}_1$ tends to exceed the population slope β_1 (upward bias); when it is negative, $\hat{\beta}_1$ tends to fall short of β_1 (downward bias). Each example names the regression, the unobserved factor in u_i, the selection mechanism and the resulting sign.

(i) **Investment gain on experience.** Take $\text{gain}_i = \beta_0 + \beta_1 \text{experience}_i + u_i$, where gain_i is investor i's annual gain (percent per year), experience_i is years of investing and u_i contains unobserved investing ability.

- Low-ability investors tend to lose money early and leave the market, so the investors observed with many years of experience are mostly high-ability ones.
- Therefore large experience_i goes together with large u_i in the observed sample: $\text{Cov}(\text{experience}_i, u_i) > 0$, so $E[u_i \mid \text{experience}_i] \neq 0$ and Least Squares Assumption 1 fails.
- By the decomposition above, $\hat{\beta}_1$ is biased upward: it overstates the population return to one more year of experience (percentage points of annual gain per year), because part of the estimated slope is the ability of the survivors.

(ii) **Mutual-fund returns on track record.** Take $\text{return}_i = \beta_0 + \beta_1 \text{record}_i + u_i$, where return_i is fund i's average annual return (percent), record_i is the number of years the fund has been operating and u_i contains unobserved manager skill.

- Funds with poor returns are closed or merged and are removed from the fund databases, so a long record is observed only for funds whose managers performed well.
- Therefore $\text{Cov}(\text{record}_i, u_i) > 0$ among the funds in the database, and Least Squares Assumption 1 fails.
- $\hat{\beta}_1$ is biased upward, and the average return of the funds in the database overstates the return an investor would have earned in a fund picked at launch.

(iii) **Online ratings on price.** Take $\text{rating}_i = \beta_0 + \beta_1 \text{price}_i + u_i$, where rating_i is a product rating (stars), price_i is the price paid (dollars) and u_i is the buyer's unobserved satisfaction.

- Only buyers with very strong experiences (very satisfied or very disappointed) take the trouble to post a rating, so the posted ratings are selected on u_i and are polarised towards one and five stars.
- If disappointed buyers of expensive items are the most likely to post (high expectations not met), the posted sample has $\text{Cov}(\text{price}_i, u_i) < 0$, Least Squares Assumption 1 fails and $\hat{\beta}_1$ is biased downward.
- The mean posted rating is a sample statistic of the reviewers only; it does not estimate the population mean satisfaction of all buyers.

Examples of selection bias: investment gain on experience (survivorship of able investors); mutual-fund returns on track record (failed funds dropped from databases); online ratings on price (only buyers with extreme experiences post).

Part (a), Method 2: Selected sample versus target population

Let $S_i = 1$ if unit i of the target population appears in the data and $S_i = 0$ otherwise. OLS on the observed data estimates the conditional mean of the *selected* units,

$$E[Y_i \,|\, X_i, S_i = 1] = \beta_0 + \beta_1 X_i + E[u_i \,|\, X_i, S_i = 1].$$

If inclusion S_i depends on u_i (or on Y_i, which contains u_i), the last term is not zero and in general changes with X_i; the slope fitted to the selected sample then mixes the population slope β_1 with the slope of $E[u_i \,|\, X_i, S_i = 1]$ in X_i. The target population is the one the economic question is about; the sample is drawn only from its selected part.

(i) **Investment gain on experience.** The target population is every investor who ever started investing; the data contain only investors still active ($S_i = 1$), and survival depends on ability u_i.

For a hypothetical numerical illustration, let the population slope be $\beta_1 = 0$ and $\beta_0 = 0$, so $\text{gain}_i = u_i$, with $u_i = +4$ percent per year for high-ability investors and $u_i = -4$ for low-ability ones. Two cohorts of 100 investors each start half high-ability and half low-ability; one is observed after 1 year, the other after 10 years, by which time 45 of its 50 low-ability investors have exited. The observed sample therefore has 100 one-year investors and 55 ten-year investors, with group means

$$\overline{\text{gain}}_1 = \frac{50(4) + 50(-4)}{100} = \frac{200 - 200}{100} = 0,$$
$$\overline{\text{gain}}_{10} = \frac{50(4) + 5(-4)}{55} = \frac{200 - 20}{55} = \frac{180}{55} = 3.27.$$

With a regressor that takes only two values $a < b$, the OLS slope is the difference in group means divided by $b - a$. To show this, let n_a, n_b be the group sizes, $n = n_a + n_b$, $p = n_b/n$ (so $n_a = n(1-p)$), and $\bar{Y}_a, \bar{Y}_b$ the group means. Then $\bar{X} = (1-p)a + pb$, so

$$X_i - \bar{X} = a - (1-p)a - pb = -p(b-a) \qquad (X_i = a),$$
$$X_i - \bar{X} = b - (1-p)a - pb = (1-p)(b-a) \qquad (X_i = b).$$

Hence

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= n_a\, p^2(b-a)^2 + n_b(1-p)^2(b-a)^2 \\ &= n(1-p)p^2(b-a)^2 + np(1-p)^2(b-a)^2 \\ &= np(1-p)(b-a)^2\big[p + (1-p)\big] = np(1-p)(b-a)^2.\end{aligned}$$

Since $\sum_{i=1}^{n}(X_i - \bar{X})\bar{Y} = \bar{Y}\sum_{i=1}^{n}(X_i - \bar{X}) = 0$,

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{n}(X_i - \bar{X})Y_i \\ &= -p(b-a)\, n_a\bar{Y}_a + (1-p)(b-a)\, n_b\bar{Y}_b \\ &= -p(b-a)\, n(1-p)\bar{Y}_a + (1-p)(b-a)\, np\,\bar{Y}_b \\ &= np(1-p)(b-a)\big(\bar{Y}_b - \bar{Y}_a\big).\end{aligned}$$

Dividing,

$$\hat{\beta}_1 = \frac{np(1-p)(b-a)(\bar{Y}_b - \bar{Y}_a)}{np(1-p)(b-a)^2} = \frac{\bar{Y}_b - \bar{Y}_a}{b-a}.$$

With $a = 1$ and $b = 10$ years,

$$\hat{\beta}_1 = \frac{3.27 - 0}{10 - 1} = \frac{180/55}{9} = 0.36,$$

that is, 0.36 percentage points of annual gain per year of experience, although the population slope is $\beta_1 = 0$. The whole estimated slope is survivorship:

$$E[u_i \,|\, \text{experience}_i = 10, S_i = 1] = 3.27 > 0 = E[u_i \,|\, \text{experience}_i = 1, S_i = 1].$$

(ii) **Mutual-fund returns on track record.** The target population is every fund ever launched; the database holds only funds still operating ($S_i = 1$), and a fund survives a long time only if its returns, and hence its u_i, were high. So $E[u_i \,|\, \text{record}_i, S_i = 1]$ rises with record_i and $\hat{\beta}_1$ is biased upward; the sample mean return of surviving funds exceeds the population mean return of all launched funds.

(iii) **Online ratings on price.** The target population is every buyer; the data hold only buyers who post ($S_i = 1$), and posting is more likely when $|u_i|$ is large. The sample mean rating estimates $E[\text{rating}_i \,|\, S_i = 1]$, not the population mean $E[\text{rating}_i]$, and the posted distribution is polarised; if disappointed buyers of expensive items post most, $E[u_i \,|\, \text{price}_i, S_i = 1]$ falls with price and $\hat{\beta}_1$ is biased downward.

Examples of selection bias: investment gain on experience (survivorship of able investors); mutual-fund returns on track record (failed funds dropped from databases); online ratings on price (only buyers with extreme experiences post).

Unofficial

Part (b), Method 1: Decomposition TSS = ESS + SSR and the ratio ESS/TSS

Let $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ be the OLS predicted value and $\hat{u}_i = Y_i - \hat{Y}_i$ the OLS residual, so that

$$Y_i = \hat{Y}_i + \hat{u}_i.$$

Define the total, explained and residual sums of squares

$$\text{TSS} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad \text{ESS} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2, \qquad \text{SSR} = \sum_{i=1}^{n}\hat{u}_i^2.$$

Subtract $\bar{Y}$ from both sides of $Y_i = \hat{Y}_i + \hat{u}_i$:

$$Y_i - \bar{Y} = (\hat{Y}_i - \bar{Y}) + \hat{u}_i.$$

Square both sides:

$$(Y_i - \bar{Y})^2 = (\hat{Y}_i - \bar{Y})^2 + 2(\hat{Y}_i - \bar{Y})\hat{u}_i + \hat{u}_i^2.$$

Sum over $i = 1, \ldots, n$:

$$\text{TSS} = \text{ESS} + 2\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})\hat{u}_i + \text{SSR}.$$

The cross term is zero. Recall the two OLS first-order conditions, $\sum_{i=1}^{n}\hat{u}_i = 0$ and $\sum_{i=1}^{n} X_i\hat{u}_i = 0$. Then

$$\sum_{i=1}^{n}\hat{Y}_i\hat{u}_i = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 X_i)\hat{u}_i = \hat{\beta}_0\sum_{i=1}^{n}\hat{u}_i + \hat{\beta}_1\sum_{i=1}^{n}X_i\hat{u}_i = \hat{\beta}_0 \cdot 0 + \hat{\beta}_1 \cdot 0 = 0,$$

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})\hat{u}_i = \sum_{i=1}^{n}\hat{Y}_i\hat{u}_i - \bar{Y}\sum_{i=1}^{n}\hat{u}_i = 0 - \bar{Y}\cdot 0 = 0.$$

Therefore

$$\text{TSS} = \text{ESS} + \text{SSR}.$$

Both ESS and SSR are sums of squares, so ESS ≥ 0 and SSR ≥ 0; hence $0 \leq \text{ESS} \leq \text{TSS}$. Dividing by TSS > 0 (the Y_i are not all equal) gives $0 \leq \text{ESS/TSS} \leq 1$. The measure of fit R^2 is defined as this ratio, and substituting ESS = TSS − SSR,

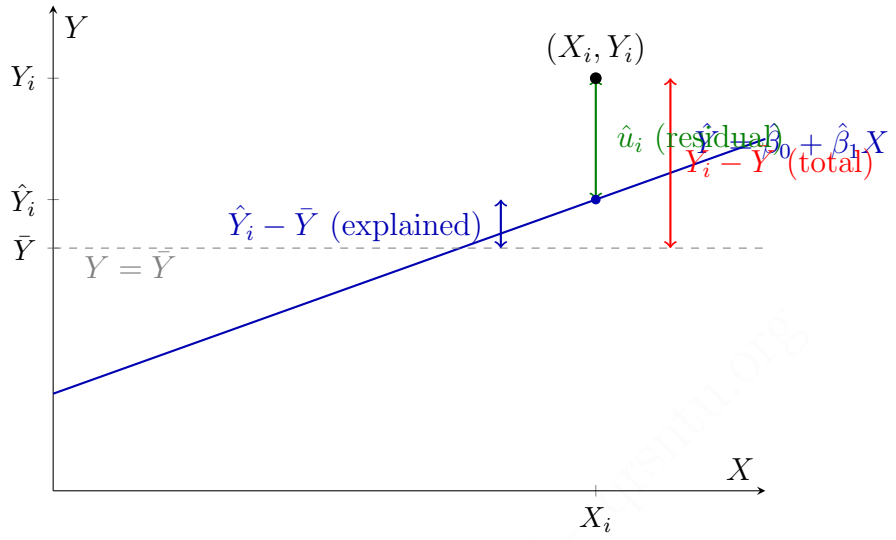
$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{TSS} - \text{SSR}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}.$$

- R^2 is the fraction of the sample variation of Y_i around $\bar{Y}$ that is explained by the OLS regression line in X_i; the remaining fraction SSR/TSS is left in the residuals.
- R^2 is unit-free: ESS and TSS are both measured in squared units of Y, so the units cancel, and rescaling Y (pounds to kilograms) or X leaves R^2 unchanged.
- $R^2 = 0$ when $\hat{\beta}_1 = 0$, because the fitted line is flat at $\bar{Y}$ and ESS $= 0$ (Question 3); $R^2 = 1$ when every $\hat{u}_i = 0$, because all points lie on the line and SSR $= 0$.

Unauthorized use is strictly prohibited.

- R^2 is a sample statistic computed from the n observations at hand, not a population parameter. A high R^2 does not show that $\hat{\beta}_1$ estimates the causal effect β_1: Least Squares Assumption 1 can fail at any value of R^2, as in Part (a).

In the diagram, for one observation (X_i, Y_i) the total deviation $Y_i - \bar{Y}$ (red) equals the explained deviation $\hat{Y}_i - \bar{Y}$ (blue) plus the residual $\hat{u}_i$ (green). Squaring and summing over all i gives $\text{TSS} = \sum_{i=1}^{n}(\text{red})^2$, $\text{ESS} = \sum_{i=1}^{n}(\text{blue})^2$ and $\text{SSR} = \sum_{i=1}^{n}(\text{green})^2$:



For a hypothetical numerical illustration take $n = 4$ observations $X = (1, 2, 3, 4)$, $Y = (2, 3, 3, 8)$. The sample means are

$$\bar{X} = \frac{1+2+3+4}{4} = \frac{10}{4} = 2.5, \qquad \bar{Y} = \frac{2+3+3+8}{4} = \frac{16}{4} = 4.$$

The deviations are $X_i - \bar{X} = (-1.5, -0.5, 0.5, 1.5)$ and $Y_i - \bar{Y} = (-2, -1, -1, 4)$, so

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) &= (-1.5)(-2) + (-0.5)(-1) + (0.5)(-1) + (1.5)(4) \\ &= 3 + 0.5 - 0.5 + 6 = 9,\end{aligned}$$

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 2.25 + 0.25 + 0.25 + 2.25 = 5.$$

Hence

$$\hat{\beta}_1 = \frac{9}{5} = 1.8, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 4 - 1.8(2.5) = 4 - 4.5 = -0.5.$$

The predicted values $\hat{Y}_i = -0.5 + 1.8X_i$ are

$$\hat{Y}_1 = -0.5 + 1.8(1) = 1.3, \qquad \hat{Y}_2 = -0.5 + 1.8(2) = 3.1,$$
$$\hat{Y}_3 = -0.5 + 1.8(3) = 4.9, \qquad \hat{Y}_4 = -0.5 + 1.8(4) = 6.7,$$

so $\hat{Y}_i - \bar{Y} = (-2.7, -0.9, 0.9, 2.7)$ and

$$\begin{aligned}\text{ESS} &= (-2.7)^2 + (-0.9)^2 + 0.9^2 + 2.7^2 = 7.29 + 0.81 + 0.81 + 7.29 = 16.20, \\ \text{TSS} &= (-2)^2 + (-1)^2 + (-1)^2 + 4^2 = 4 + 1 + 1 + 16 = 22.00, \\ \text{SSR} &= \text{TSS} - \text{ESS} = 22.00 - 16.20 = 5.80.\end{aligned}$$

In this sample 73.64% of the variation of Y around its mean is explained by the regression on X:

$$\boxed{R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}.}$$

$$\boxed{R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{16.20}{22.00} = 0.7364.}$$

Unofficial - QRS@NTU - qrsntu.org

Part (b), Method 2: Closed-form estimators and $1 - \text{SSR/TSS}$

Write

$$S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}), \qquad S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \text{TSS},$$

and use the closed-form OLS estimators

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}.$$

Substitute $\hat{\beta}_0$ into the predicted value:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i = \bar{Y} - \hat{\beta}_1\bar{X} + \hat{\beta}_1 X_i = \bar{Y} + \hat{\beta}_1(X_i - \bar{X}),$$

so $\hat{Y}_i - \bar{Y} = \hat{\beta}_1(X_i - \bar{X})$ and

$$\text{ESS} = \sum_{i=1}^{n}\hat{\beta}_1^2(X_i - \bar{X})^2 = \hat{\beta}_1^2 S_{XX}.$$

The residual is

$$\hat{u}_i = Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X}).$$

Square and sum:

$$\begin{aligned}
\text{SSR} &= \sum_{i=1}^{n}\left[(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})\right]^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - 2\hat{\beta}_1\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) + \hat{\beta}_1^2\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= S_{YY} - 2\hat{\beta}_1 S_{XY} + \hat{\beta}_1^2 S_{XX}.
\end{aligned}$$

The slope formula gives $S_{XY} = \hat{\beta}_1 S_{XX}$; substitute it into the middle term:

$$\text{SSR} = S_{YY} - 2\hat{\beta}_1^2 S_{XX} + \hat{\beta}_1^2 S_{XX} = S_{YY} - \hat{\beta}_1^2 S_{XX} = \text{TSS} - \text{ESS}.$$

So TSS = ESS + SSR follows from the closed-form slope alone, without the orthogonality of residuals. Since $\text{SSR} \geq 0$, $0 \leq \text{ESS} \leq \text{TSS}$, and for $\text{TSS} > 0$

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\text{TSS} - \text{SSR}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}.$$

Read this way, R^2 is one minus the share of the sample variation of Y that the regression line leaves unexplained.

For the same data, $S_{XX} = 5$, $S_{XY} = 9$ and $S_{YY} = 4{+}1{+}1{+}16 = 22$, so $\hat{\beta}_1 = 9/5 = 1.8$ and $\hat{\beta}_0 = 4 - 1.8(2.5) = -0.5$, with predicted values $(1.3, 3.1, 4.9, 6.7)$. The residuals are

$$\hat{u}_1 = 2 - 1.3 = 0.7, \quad \hat{u}_2 = 3 - 3.1 = -0.1, \quad \hat{u}_3 = 3 - 4.9 = -1.9, \quad \hat{u}_4 = 8 - 6.7 = 1.3,$$

so

$$\begin{aligned}
\text{SSR} &= 0.7^2 + (-0.1)^2 + (-1.9)^2 + 1.3^2 = 0.49 + 0.01 + 3.61 + 1.69 = 5.80, \\
\text{ESS} &= \hat{\beta}_1^2 S_{XX} = 1.8^2 \times 5 = 3.24 \times 5 = 16.20, \\
R^2 &= 1 - \frac{\text{SSR}}{\text{TSS}} = 1 - \frac{5.80}{22.00} = \frac{22.00 - 5.80}{22.00} = \frac{16.20}{22.00} = 0.7364.
\end{aligned}$$

$$\boxed{R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}}.}$$

$$\boxed{R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{16.20}{22.00} = 0.7364.}$$

Unofficial - QRS@NTU - qrsntu.org

Part (b), Method 3: Squared sample correlation

With S_{XX}, S_{XY}, S_{YY} as in Method 2, let r_{XY} be the sample correlation between X and Y,

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}} \qquad (S_{XX} > 0,\ S_{YY} > 0),$$

where the $(n-1)$ divisors of the sample covariance and of the two sample variances cancel. It is the sample counterpart of the population correlation $\mathrm{Corr}(X_i, Y_i)$. Using $\hat{\beta}_1 = S_{XY}/S_{XX}$ and $\hat{Y}_i - \bar{Y} = \hat{\beta}_1(X_i - \bar{X})$ (from $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$),

$$\mathrm{ESS} = \hat{\beta}_1^2 S_{XX} = \frac{S_{XY}^2}{S_{XX}^2}\, S_{XX} = \frac{S_{XY}^2}{S_{XX}} = S_{YY} \cdot \frac{S_{XY}^2}{S_{XX}S_{YY}} = r_{XY}^2\, \mathrm{TSS}.$$

Expanding $\mathrm{SSR} = \sum_{i=1}^{n}\big[(Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})\big]^2$ and substituting $\hat{\beta}_1 = S_{XY}/S_{XX}$,

$$\begin{aligned}
\mathrm{SSR} &= S_{YY} - 2\hat{\beta}_1 S_{XY} + \hat{\beta}_1^2 S_{XX} \\
&= S_{YY} - 2\,\frac{S_{XY}^2}{S_{XX}} + \frac{S_{XY}^2}{S_{XX}} \\
&= S_{YY} - \frac{S_{XY}^2}{S_{XX}} = S_{YY}\Big(1 - \frac{S_{XY}^2}{S_{XX}S_{YY}}\Big) = (1 - r_{XY}^2)\,\mathrm{TSS}.
\end{aligned}$$

Divide each by $\mathrm{TSS} > 0$:

$$\frac{\mathrm{ESS}}{\mathrm{TSS}} = r_{XY}^2, \qquad 1 - \frac{\mathrm{SSR}}{\mathrm{TSS}} = 1 - (1 - r_{XY}^2) = r_{XY}^2.$$

Both expressions therefore equal the squared sample correlation, and since $-1 \le r_{XY} \le 1$ they lie in $[0, 1]$. In this reading R^2 measures the strength of the linear association between X and Y in the sample, whatever its sign: $r_{XY} = 0.9$ and $r_{XY} = -0.9$ both give $R^2 = 0.81$.

For the same data, $S_{XY} = 9$, $S_{XX} = 5$, $S_{YY} = 22$, so

$$\begin{aligned}
r_{XY} &= \frac{9}{\sqrt{5 \times 22}} = \frac{9}{\sqrt{110}} = \frac{9}{10.4881} = 0.8581, \\
r_{XY}^2 &= \frac{81}{110} = 0.7364, \\
\mathrm{ESS} &= r_{XY}^2\,\mathrm{TSS} = \frac{81}{110} \times 22 = \frac{81}{5} = 16.20, \\
\mathrm{SSR} &= (1 - r_{XY}^2)\,\mathrm{TSS} = \frac{29}{110} \times 22 = \frac{29}{5} = 5.80.
\end{aligned}$$

$$\boxed{R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{SSR}}{\mathrm{TSS}}.}$$

$$\boxed{R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{16.20}{22.00} = 0.7364.}$$

Question 2 (Predicting weight from height)

Problem

A random sample of 100 20-year-old men is selected from a population and these men's height and weight are recorded. An OLS regression of weight on height yields

$$\widehat{\text{Weight}} = -79.24 + 4.16 \times \text{Height}, \tag{1}$$

where Weight is measured in pounds and Height is measured in inches.

(a) What is the regression's weight prediction for someone who is 64 inches tall? 68 inches tall? 72 inches tall?

(b) A man has a late growth spurt and grows 2 inches over the course of a year. What is the regression's prediction for the increase in this man's weight?

(Tutorial Question Set No. 4, Q2. No official solution is given on the Tutorial 4 slides – the "Question 2" slide is blank – so this solution is supplied in full.)

Solution

Method 1: Direct substitution into the sample regression line

(a) Substitute each height (inches) into equation (1); compute the product $4.16 \times \text{Height}$ first and then add the intercept -79.24:

$$\begin{aligned}
\widehat{\text{Weight}}(64) &= -79.24 + 4.16(64) = -79.24 + 266.24 = 187.00, \\
\widehat{\text{Weight}}(68) &= -79.24 + 4.16(68) = -79.24 + 282.88 = 203.64, \\
\widehat{\text{Weight}}(72) &= -79.24 + 4.16(72) = -79.24 + 299.52 = 220.28.
\end{aligned}$$

- Each value is the estimated average weight, in pounds, of 20-year-old men of that height. It is read off the sample regression line fitted to the 100 sampled men, which estimates the population regression line $\beta_0 + \beta_1 \text{Height}$.
- An individual man who is 68 inches tall will in general weigh more or less than 203.64 pounds; the gap is his error u_i, which contains everything other than height that affects weight (diet, exercise, build).
- The intercept -79.24 is the prediction at a height of 0 inches, far outside the heights of adult men; it fixes the level of the line and has no meaning of its own.

$$\boxed{\begin{aligned}
&\widehat{\text{Weight}}(64) = 187.00, \\
&\widehat{\text{Weight}}(68) = 203.64, \\
&\widehat{\text{Weight}}(72) = 220.28 \text{ pounds.}
\end{aligned}}$$

(b) Let the man be 64 inches tall before the growth spurt, so he is $64 + 2 = 66$ inches tall after it. The prediction after the spurt is

$$\widehat{\text{Weight}}(66) = -79.24 + 4.16(66) = -79.24 + 274.56 = 195.32,$$

and the prediction before it is $\widehat{\text{Weight}}(64) = 187.00$ from part (a). The predicted increase is

$$\widehat{\text{Weight}}(66) - \widehat{\text{Weight}}(64) = 195.32 - 187.00 = 8.32 = 4.16 \times 2.$$

- The regression predicts that the man's weight rises by 8.32 pounds: 4.16 pounds per inch times 2 inches.
- The slope $\hat{\beta}_1 = 4.16$ is estimated by comparing different men of different heights in one cross-section, and it estimates the population slope β_1. Applying it to one man's growth over a year assumes that a within-person change in height moves expected weight in the same way as the cross-sectional comparison.
- The 8.32 pounds is a change in his predicted (expected) weight; his actual change also depends on diet and exercise over the year, which sit in u_i.

$$\boxed{\Delta\widehat{\text{Weight}} = \hat{\beta}_1 \times 2 = 8.32 \text{ pounds.}}$$

Unofficial QRS@NTU - qrsntu.org

Method 2: Slope times change in height

For any two heights h and $h + \Delta h$ (inches), subtract the two predictions from equation (1):

$$\begin{aligned}\widehat{\text{Weight}}(h + \Delta h) - \widehat{\text{Weight}}(h) &= \big[-79.24 + 4.16(h + \Delta h)\big] - \big[-79.24 + 4.16h\big] \\ &= -79.24 + 4.16h + 4.16\,\Delta h + 79.24 - 4.16h \\ &= 4.16\,\Delta h = \hat{\beta}_1\,\Delta h.\end{aligned}$$

The intercept cancels, so the change in the prediction depends only on the change in height, at $\hat{\beta}_1 = 4.16$ pounds per inch, whatever the starting height.

(a) Anchor the line at the smallest height,

$$\widehat{\text{Weight}}(64) = -79.24 + 4.16(64) = -79.24 + 266.24 = 187.00.$$

The heights 64, 68, 72 are $\Delta h = 4$ inches apart, and each 4-inch step adds

$$\hat{\beta}_1\,\Delta h = 4.16 \times 4 = 16.64 \text{ pounds.}$$

Hence

$$\begin{aligned}\widehat{\text{Weight}}(68) &= \widehat{\text{Weight}}(64) + 16.64 = 187.00 + 16.64 = 203.64, \\ \widehat{\text{Weight}}(72) &= \widehat{\text{Weight}}(68) + 16.64 = 203.64 + 16.64 = 220.28.\end{aligned}$$

The predictions are estimated average weights of 20-year-old men of each height, rising by a constant 16.64 pounds per 4 inches because the fitted relation is linear.

$$\boxed{\begin{aligned}&\widehat{\text{Weight}}(64) = 187.00, \\ &\widehat{\text{Weight}}(68) = 203.64, \\ &\widehat{\text{Weight}}(72) = 220.28 \text{ pounds.}\end{aligned}}$$

(b) The growth spurt is $\Delta h = 2$ inches, so by the general difference above, for any starting height,

$$\Delta\widehat{\text{Weight}} = \hat{\beta}_1\,\Delta h = 4.16 \times 2 = 8.32 \text{ pounds.}$$

This is the predicted change in expected weight; it treats the cross-sectional slope as the effect of one more inch of height for this man.

$$\boxed{\Delta\widehat{\text{Weight}} = \hat{\beta}_1 \times 2 = 8.32 \text{ pounds.}}$$

Question 3 (Consequences of $\hat{\beta}_1 = 0$ for $\hat{Y}_i$ and R^2)

Problem

(a) Show that the average of the OLS predicted values $\hat{Y}_i$, $\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i$, is equal to the sample average of Y, $\bar{Y}$.

(b) Show that if the OLS estimator $\hat{\beta}_1 = 0$, then $\hat{Y}_i = \hat{\beta}_0 = \bar{Y}$.

(c) Show that if (b) holds, then $R^2 = 0$.

(Tutorial Question Set No. 4, Q3.)

Solution

Method 1: First-order conditions

The OLS estimators $\hat{\beta}_0, \hat{\beta}_1$ minimise $S(b_0, b_1) = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$. Differentiating with respect to b_0 by the chain rule (inner derivative -1),

$$\frac{\partial S}{\partial b_0} = \sum_{i=1}^{n} 2\big(Y_i - b_0 - b_1 X_i\big)(-1) = -2\sum_{i=1}^{n}\big(Y_i - b_0 - b_1 X_i\big).$$

At the minimiser this is zero; dividing by -2 gives the first-order condition for the intercept,

$$\sum_{i=1}^{n}\big(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\big) = 0, \qquad \text{that is,} \qquad \sum_{i=1}^{n}\hat{u}_i = 0.$$

(a) By definition of the residual, $\hat{u}_i = Y_i - \hat{Y}_i$, so $Y_i = \hat{Y}_i + \hat{u}_i$. Sum over i and use the first-order condition:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}\hat{Y}_i + \sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}\hat{Y}_i + 0 = \sum_{i=1}^{n}\hat{Y}_i.$$

Divide by n:

$$\boxed{\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}.}$$

(b) Set $\hat{\beta}_1 = 0$ in the first-order condition for the intercept:

$$\sum_{i=1}^{n}\big(Y_i - \hat{\beta}_0 - 0 \cdot X_i\big) = \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 = 0 \quad \Longrightarrow \quad \hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}.$$

The predicted value of every observation is then

$$\hat{Y}_i = \hat{\beta}_0 + 0 \cdot X_i = \hat{\beta}_0 = \bar{Y}.$$

$$\boxed{\hat{Y}_i = \hat{\beta}_0 = \bar{Y} \quad \text{for every } i = 1, \ldots, n.}$$

(c) By part (b), $\hat{Y}_i - \bar{Y} = \bar{Y} - \bar{Y} = 0$ for every i, so

$$\text{ESS} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} 0^2 = 0.$$

With $\text{TSS} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 > 0$ (the Y_i are not all equal), the measure of fit is $R^2 = \text{ESS}/\text{TSS}$. The fitted line is flat at the sample mean $\bar{Y}$, so X explains none of the sample variation of Y.

$$\boxed{R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{0}{\text{TSS}} = 0 \qquad (\text{provided TSS} > 0).}$$

Unofficial QRS@NTU - qrsntu.org

Method 2: Closed-form estimators

Use the closed-form OLS estimators

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X},$$

together with $\sum_{i=1}^{n} Y_i = n\bar{Y}$ and $\sum_{i=1}^{n} X_i = n\bar{X}$.

(a) Sum the predicted values $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$:

$$\begin{aligned}\sum_{i=1}^{n} \hat{Y}_i &= \sum_{i=1}^{n}\big(\hat{\beta}_0 + \hat{\beta}_1 X_i\big) = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = n\hat{\beta}_0 + n\hat{\beta}_1 \bar{X} \\ &= n\big(\bar{Y} - \hat{\beta}_1 \bar{X}\big) + n\hat{\beta}_1 \bar{X} = n\bar{Y} - n\hat{\beta}_1 \bar{X} + n\hat{\beta}_1 \bar{X} = n\bar{Y}.\end{aligned}$$

Divide by n:

$$\boxed{\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}.}$$

(b) Substitute $\hat{\beta}_1 = 0$ into the intercept formula and into the predicted value:

$$\hat{\beta}_0 = \bar{Y} - 0 \cdot \bar{X} = \bar{Y}, \qquad \hat{Y}_i = \hat{\beta}_0 + 0 \cdot X_i = \hat{\beta}_0 = \bar{Y}.$$

$$\boxed{\hat{Y}_i = \hat{\beta}_0 = \bar{Y} \quad \text{for every } i = 1, \ldots, n.}$$

(c) With $\hat{\beta}_1 = 0$ and $\hat{\beta}_0 = \bar{Y}$ the residual is

$$\hat{u}_i = Y_i - \hat{\beta}_0 - 0 \cdot X_i = Y_i - \bar{Y},$$

so

$$\mathrm{SSR} = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \mathrm{TSS}.$$

From the decomposition $\mathrm{TSS} = \mathrm{ESS} + \mathrm{SSR}$ (Question 1(b)),

$$\mathrm{ESS} = \mathrm{TSS} - \mathrm{SSR} = \mathrm{TSS} - \mathrm{TSS} = 0,$$

and equivalently $1 - \mathrm{SSR}/\mathrm{TSS} = 1 - 1 = 0$: every unit of sample variation in Y is left in the residuals.

$$\boxed{R^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{0}{\mathrm{TSS}} = 0 \qquad (\text{provided } \mathrm{TSS} > 0).}$$

Method 3: Deviation form, part (a) and sample correlation

(a) Substitute $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ into the predicted value and subtract $\bar{Y}$:

$$\hat{Y}_i - \bar{Y} = \left(\bar{Y} - \hat{\beta}_1 \bar{X}\right) + \hat{\beta}_1 X_i - \bar{Y} = \hat{\beta}_1 (X_i - \bar{X}).$$

Sum over i and use $\sum_{i=1}^{n} (X_i - \bar{X}) = 0$:

$$\sum_{i=1}^{n} (X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0,$$
$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}) = \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X}) = \hat{\beta}_1 \cdot 0 = 0.$$

Hence $\sum_{i=1}^{n} \hat{Y}_i - n\bar{Y} = 0$, so $\sum_{i=1}^{n} \hat{Y}_i = n\bar{Y}$. Divide by n:

$$\boxed{\frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n} \sum_{i=1}^{n} Y_i = \bar{Y}.}$$

(b) If $\hat{\beta}_1 = 0$, then $\hat{Y}_i = \hat{\beta}_0 + 0 \cdot X_i = \hat{\beta}_0$ for every i: all predicted values equal the same constant $\hat{\beta}_0$. Their average is therefore

$$\frac{1}{n} \sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n} \sum_{i=1}^{n} \hat{\beta}_0 = \frac{1}{n} \cdot n\hat{\beta}_0 = \hat{\beta}_0.$$

By part (a) the same average equals $\bar{Y}$. Equating the two expressions gives $\hat{\beta}_0 = \bar{Y}$.

$$\boxed{\hat{Y}_i = \hat{\beta}_0 = \bar{Y} \quad \text{for every } i = 1, \ldots, n.}$$

(c) By Question 1(b), Method 3, $R^2 = r_{XY}^2$ with

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}, \qquad S_{XX} = \sum_{i=1}^{n} (X_i - \bar{X})^2 > 0, \quad S_{YY} = \text{TSS} > 0.$$

The slope formula $\hat{\beta}_1 = S_{XY}/S_{XX}$ with $\hat{\beta}_1 = 0$ and $S_{XX} > 0$ gives

$$S_{XY} = \hat{\beta}_1 S_{XX} = 0 \cdot S_{XX} = 0,$$

so

$$r_{XY} = \frac{0}{\sqrt{S_{XX} S_{YY}}} = 0, \qquad R^2 = r_{XY}^2 = 0.$$

A zero OLS slope is the same event as a zero sample correlation between X and Y, and then $\text{ESS} = r_{XY}^2 \text{TSS} = 0$.

$$\boxed{R^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{0}{\text{TSS}} = 0 \qquad (\text{provided TSS} > 0).}$$

Question 4 (CAPM: comparing assets with equal β but different α)

Problem

As explained in class, we can use OLS to estimate the Capital Asset Pricing Model (CAPM). The linear regression model is defined as

$$(R - R_f) = \alpha + \beta(R_m - R_f) + u, \tag{2}$$

where $R - R_f$ denotes the excess return of the stock (or hedge fund), $R_m - R_f$ denotes the excess return of the market (e.g., R_m can be the return of S&P 500), and R_f denotes the return of the risk-free asset (e.g., US government bond). In practice, we report the OLS estimates $\hat{\alpha}$ and $\hat{\beta}$ for the stock (or fund). Under the standard CAPM theory, $\alpha = 0$. Therefore, if we find $\hat{\alpha}$ is significantly positive for a certain stock (or fund), then it is especially valuable as it can beat the prediction from CAPM (i.e., it has a premium that cannot be explained by CAPM). Compare the return and risk of a stock (or a fund) with different values of α but the same value of β. What do you find? [Hint: find its expected value and variance.] *(Tutorial Question Set No. 4, Q4. No official solution is given on the Tutorial 4 slides – both "Question 4" slides in the deck are blank – so this solution is supplied in full, extending the Lecture 4 CAPM material.)*

Solution

Throughout, write $Y = R - R_f$ for the asset's excess return and $X = R_m - R_f$ for the market's excess return, so that equation (2) reads $Y = \alpha + \beta X + u$ with population parameters α, β. Compare two funds A and B with

$$Y_A = \alpha_A + \beta X + u_A, \qquad Y_B = \alpha_B + \beta X + u_B, \qquad \alpha_B > \alpha_A, \qquad \mathrm{Var}(u_A) = \mathrm{Var}(u_B) = \sigma_u^2,$$

the same β and the same idiosyncratic risk, and with Least Squares Assumption 1 in population form, $E[u_A \mid X] = E[u_B \mid X] = 0$. Below, u stands for either u_A or u_B.

Method 1: Expectation and variance rules

By the law of iterated expectations, Least Squares Assumption 1 gives

$$\begin{aligned} E[u] &= E\big[E[u \mid X]\big] = E[0] = 0, \\ \mathrm{Cov}(X, u) &= E[Xu] - E[X]\,E[u] = E\big[X\,E[u \mid X]\big] - E[X] \cdot 0 = E[X \cdot 0] = 0. \end{aligned}$$

Return. By linearity of expectation, with α and β constants,

$$E[R - R_f] = E[\alpha + \beta X + u] = \alpha + \beta\,E[X] + E[u] = \alpha + \beta\,E[R_m - R_f].$$

For the two funds,

$$E[R_B - R_f] - E[R_A - R_f] = \big(\alpha_B + \beta\,E[X]\big) - \big(\alpha_A + \beta\,E[X]\big) = \alpha_B - \alpha_A > 0.$$

Risk. Adding a constant does not change a variance, and $\mathrm{Var}(aZ+W) = a^2\,\mathrm{Var}(Z) + \mathrm{Var}(W) + 2a\,\mathrm{Cov}(Z,W)$ for a constant a. Hence

$$\begin{aligned}\mathrm{Var}(R-R_f) &= \mathrm{Var}(\alpha+\beta X+u) = \mathrm{Var}(\beta X+u)\\ &= \beta^2\,\mathrm{Var}(X) + \mathrm{Var}(u) + 2\beta\,\mathrm{Cov}(X,u)\\ &= \beta^2\,\mathrm{Var}(R_m-R_f) + \mathrm{Var}(u) + 2\beta\cdot 0 = \beta^2\,\mathrm{Var}(R_m-R_f) + \mathrm{Var}(u).\end{aligned}$$

For the two funds,

$$\mathrm{Var}(R_B-R_f) - \mathrm{Var}(R_A-R_f) = \left[\beta^2\,\mathrm{Var}(X)+\sigma_u^2\right] - \left[\beta^2\,\mathrm{Var}(X)+\sigma_u^2\right] = 0.$$

- α shifts the expected excess return one for one: fund B earns $\alpha_B-\alpha_A$ more per period on average, in the same units as the returns (percent per month in the illustration below).
- α does not enter the variance. Risk is $\beta^2\,\mathrm{Var}(R_m-R_f)$ (systematic or market risk, set by β) plus $\mathrm{Var}(u)$ (idiosyncratic risk), and both pieces are equal for the two funds.
- With the same β and the same $\mathrm{Var}(u)$, the fund with the higher α therefore offers a higher expected return for the same risk. Under the CAPM $\alpha = 0$, so a fund with $\alpha > 0$ earns a premium that the CAPM does not explain.
- α, β and $\mathrm{Var}(u)$ are population parameters. In practice they are estimated by OLS from a sample of returns, and a positive $\hat{\alpha}$ is evidence of a positive α only if it is statistically significant.

For a hypothetical numerical illustration take monthly data with $E[R_m-R_f] = 0.6\%$, $\mathrm{sd}(R_m-R_f) = 4\%$, $\beta = 1.2$ and $\mathrm{sd}(u) = 2\%$ for both funds, and $\alpha_A = 0$, $\alpha_B = 0.3\%$. Then

$$\begin{aligned}\mathrm{Var}(R-R_f) &= (1.2)^2(0.04)^2 + (0.02)^2 = 1.44\times 0.0016 + 0.0004\\ &= 0.002304 + 0.0004 = 0.002704,\\ \mathrm{sd}(R-R_f) &= \sqrt{0.002704} = 0.0520 = 5.20\% \quad \text{(both funds)},\\ E[R_A-R_f] &= 0 + 1.2(0.6\%) = 0.72\%,\\ E[R_B-R_f] &= 0.3\% + 1.2(0.6\%) = 0.3\% + 0.72\% = 1.02\%.\end{aligned}$$

The gap is $E[R_B-R_f] - E[R_A-R_f] = 1.02\% - 0.72\% = 0.30$ percentage points per month.

$$\boxed{\mathrm{Var}(R-R_f) = \beta^2\,\mathrm{Var}(R_m-R_f) + \mathrm{Var}(u),}$$

$$\boxed{\text{Fund } B \text{ earns 0.30 pp more per month, at the identical risk sd} = 5.20\%.}$$

Method 2: Variance from the definition $E\big[(Y - E[Y])^2\big]$

Return. Taking expectations of $Y = \alpha + \beta X + u$ and using $E[u] = E\big[E[u \mid X]\big] = 0$,

$$E[Y] = \alpha + \beta\, E[X] + E[u] = \alpha + \beta\, E[X],$$

so $E[Y_B] - E[Y_A] = \alpha_B - \alpha_A > 0$.

Risk. The deviation of Y from its mean is

$$Y - E[Y] = \big(\alpha + \beta X + u\big) - \big(\alpha + \beta\, E[X]\big) = \beta\big(X - E[X]\big) + u.$$

The constant α cancels at this step, so it cannot appear in the variance. Square the deviation:

$$\big(Y - E[Y]\big)^2 = \beta^2\big(X - E[X]\big)^2 + 2\beta\big(X - E[X]\big)u + u^2.$$

Take expectations term by term:

$$\mathrm{Var}(Y) = E\big[(Y - E[Y])^2\big] = \beta^2\, E\big[(X - E[X])^2\big] + 2\beta\, E\big[(X - E[X])u\big] + E[u^2].$$

Evaluate each expectation. The first is the definition of $\mathrm{Var}(X)$. For the second, $X - E[X]$ is a function of X, so the law of iterated expectations gives

$$E\big[(X - E[X])u\big] = E\big[(X - E[X])\, E[u \mid X]\big] = E\big[(X - E[X]) \cdot 0\big] = 0.$$

For the third,

$$E[u^2] = \mathrm{Var}(u) + \big(E[u]\big)^2 = \mathrm{Var}(u) + 0^2 = \mathrm{Var}(u).$$

Therefore

$$\mathrm{Var}(Y) = \beta^2\, \mathrm{Var}(X) + 2\beta \cdot 0 + \mathrm{Var}(u) = \beta^2\, \mathrm{Var}(R_m - R_f) + \mathrm{Var}(u),$$

the same for both funds because they share β and σ_u^2.

With the monthly numbers of Method 1,

$$\begin{aligned} E\big[(Y - E[Y])^2\big] &= (1.2)^2(0.04)^2 + 2(1.2)(0) + (0.02)^2 \\ &= 0.002304 + 0 + 0.0004 = 0.002704, \\ \mathrm{sd}(Y) &= \sqrt{0.002704} = 0.0520 = 5.20\% \quad \text{(both funds)}, \\ E[Y_A] &= 0 + 1.2(0.6\%) = 0.72\%, \qquad E[Y_B] = 0.3\% + 0.72\% = 1.02\%, \end{aligned}$$

a gap of $1.02\% - 0.72\% = 0.30$ percentage points per month.

$$\boxed{\mathrm{Var}(R - R_f) = \beta^2\, \mathrm{Var}(R_m - R_f) + \mathrm{Var}(u),}$$

$$\boxed{\text{Fund } B \text{ earns 0.30 pp more per month, at the identical risk sd} = 5.20\%.}$$

Method 3: Conditional mean and the law of total variance

Note: uses the law of total variance, beyond the Lecture 4 scope.

Condition on the market excess return X. Since α, β are constants, $\alpha + \beta X$ is fixed once X is known, and $E[u \mid X] = 0$:

$$E[Y \mid X] = \alpha + \beta X + E[u \mid X] = \alpha + \beta X,$$
$$\mathrm{Var}(Y \mid X) = \mathrm{Var}(\alpha + \beta X + u \mid X) = \mathrm{Var}(u \mid X).$$

At every market outcome $X = x$ the two funds' conditional expected excess returns differ by

$$E[Y_B \mid X = x] - E[Y_A \mid X = x] = (\alpha_B + \beta x) - (\alpha_A + \beta x) = \alpha_B - \alpha_A,$$

so fund B is ahead by the same margin in rising and in falling markets.

Return. By the law of iterated expectations,

$$E[Y] = E\big[E[Y \mid X]\big] = E[\alpha + \beta X] = \alpha + \beta\, E[R_m - R_f].$$

Risk. The law of total variance states $\mathrm{Var}(Y) = \mathrm{Var}\big(E[Y \mid X]\big) + E\big[\mathrm{Var}(Y \mid X)\big]$. Its two pieces are

$$\mathrm{Var}\big(E[Y \mid X]\big) = \mathrm{Var}(\alpha + \beta X) = \beta^2\, \mathrm{Var}(X), \qquad E\big[\mathrm{Var}(Y \mid X)\big] = E\big[\mathrm{Var}(u \mid X)\big].$$

Apply the same law to u:

$$\mathrm{Var}(u) = \mathrm{Var}\big(E[u \mid X]\big) + E\big[\mathrm{Var}(u \mid X)\big] = \mathrm{Var}(0) + E\big[\mathrm{Var}(u \mid X)\big] = E\big[\mathrm{Var}(u \mid X)\big].$$

Therefore

$$\mathrm{Var}(Y) = \beta^2\, \mathrm{Var}(R_m - R_f) + \mathrm{Var}(u).$$

The first piece is the variance of the conditional mean line $\alpha + \beta X$; shifting that line up by α moves every point by the same constant and leaves its variance unchanged.

With the monthly numbers of Method 1, in a month with $x = -4\%$,

$$E[Y_A \mid X = -4\%] = 0 + 1.2(-4\%) = -4.80\%,$$
$$E[Y_B \mid X = -4\%] = 0.3\% + 1.2(-4\%) = 0.3\% - 4.80\% = -4.50\%,$$

and unconditionally

$$\mathrm{Var}\big(E[Y \mid X]\big) = (1.2)^2(0.04)^2 = 1.44 \times 0.0016 = 0.002304,$$
$$E\big[\mathrm{Var}(Y \mid X)\big] = \mathrm{Var}(u) = (0.02)^2 = 0.0004,$$
$$\mathrm{Var}(Y) = 0.002304 + 0.0004 = 0.002704,$$
$$\mathrm{sd}(Y) = \sqrt{0.002704} = 0.0520 = 5.20\% \quad \text{(both funds)},$$
$$E[Y_A] = 0 + 1.2(0.6\%) = 0.72\%,$$
$$E[Y_B] = 0.3\% + 1.2(0.6\%) = 1.02\%,$$

a gap of 0.30 percentage points per month.

$$\boxed{\mathrm{Var}(R - R_f) = \beta^2\, \mathrm{Var}(R_m - R_f) + \mathrm{Var}(u),}$$

Fund B earns 0.30 pp more per month, at the identical risk sd = 5.20%.

Final Answers

1(a) Investment gain on experience; mutual-fund returns on track record; online ratings on price.

1(b) $R^2 = \text{ESS}/\text{TSS} = 1 - \text{SSR}/\text{TSS}$; example: TSS $= 22.00$, ESS $= 16.20$, SSR $= 5.80$, $R^2 = 0.7364$.

2(a) $\widehat{\text{Weight}}(64) = 187.00$, $\widehat{\text{Weight}}(68) = 203.64$, $\widehat{\text{Weight}}(72) = 220.28$ pounds.

2(b) $\Delta\widehat{\text{Weight}} = \hat{\beta}_1 \times 2 = 8.32$ pounds.

3(a) $\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \bar{Y}$.

3(b) $\hat{Y}_i = \hat{\beta}_0 = \bar{Y}$ for every i.

3(c) $\text{ESS} = 0 \Rightarrow R^2 = 0$ (TSS > 0).

4 $E[R - R_f] = \alpha + \beta E[R_m - R_f]$; $\text{Var}(R - R_f) = \beta^2 \,\text{Var}(R_m - R_f) + \text{Var}(u)$; Fund B earns 0.30 pp more per month at identical sd $= 5.20\%$.

Unofficial QRS@NTU answer set