Module 11: Statistical inference for simple linear regression practice problems

(The attached PDF file has better formatting.)

** Exercise 11.1: Key assumptions of classical regression analysis

Classical regression analysis is based on the statistical model $y_j = \alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk} + \epsilon_j$.

Explain the following key assumptions of classical regression analysis:

A. Linearity of the relation between the explanatory variables and the response variable
B. Constant variance of the error term
C. Normality of the error term
D. Independence of the error terms
E. Explanatory variables are fixed or have no measurement error
F. X values are not invariant: they have at least two values

Part A: The linearity assumption is that the expected error term is zero: $E(\epsilon_j) = 0$

Jacob: Do you mean that the average y-value is a linear function of the average x-values? We show this by taking expectations:

$$E(y_j) = E(\alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk} + \epsilon_j) \Rightarrow \overline{y} = \alpha + \beta_1 \overline{x}_1 + \beta_2 \overline{x}_2 + \ldots + \beta_k \overline{x}_k$$

Rachel: Your explanation shows that the means of the observed X values and the mean of the observed Y values lie on the fitted regression line. The value of A (the least squares estimator of α) is chosen to force this relation. Mean squared estimators have the property that

$$\overline{y} = A + B_1 \overline{x}_1 + B_2 \overline{x}_2 + \ldots + B_k \overline{x}_k$$

This relation is true even if the true relation between Y and the X-values is not linear. It reflects the estimation method, not the attributes of the relation.

The linearity assumption is that the expected error term is zero at each point: $E(\epsilon_j) = 0$ for all *j*.

Jacob: Are you saying that the expected value of each observed residual is zero?

Rachel: Distinguish between the observed residual and the error term.

- The observed residual is the observed response variable Y minus the linear function of the observed explanatory variables, using the ordinary least squares estimators (A and B) for the parameters.
- The error term is the expected response variable Y at any point minus the linear function of the observed explanatory variables, using the population regression parameters α and β.

The population regression parameters α and β specify the relation between the explanatory variables and the response variable. The ordinary least squares estimators A and B are linear functions of the sample values.

A and B are not the same as α and β. The linearity assumption says the expected value of the error term at each point is zero, not that the expected value of the residual using the ordinary least squares estimators is zero at each point.

Jacob: Can you show how the linearity assumption relates to the expected value of the error term?

Rachel: Suppose the true relation is $Y = 1 + X^2 + \epsilon$, which is not linear. We fit a regression line through three points: X = (0, 1, 2), for which the expected Y values are (1, 2, 5).

The ordinary least squares estimators A and B depend on the sample values. They are unbiased estimators of α and β. They are linear functions of the observed y-values, so the expected values of A and B are fitted values using the expected values of Y at each point. This fitted regression line is Y = 0.66667 + 2X + ϵ.

The expected residuals are computed in the table below:

X-value	*Expected Y Value*	*Fitted Y Value*	*Expected Residual*
0	1	0.66667	0.33333
1	2	2.66667	-0.66667
2	5	4.66667	0.33333

- X = 0: expected residual = 0.33333.
- X = 1: expected residual = –0.66667.
- X = 2: expected residual = 0.33333.

Intuition: The linearity assumption says that the expected error term *at each X value* is zero, not that the mean residuals for the sample points is zero. The second statement is an attribute of least squares estimators; the first statement is an assumption of classical regression analysis.

Jacob: Is the expected error term over all X values zero even if the true relation is not linear?

Rachel: The expected *residual* over all X values *is zero* whether or not the true relation is linear.

But the X values are fixed in experimental studies and are measured without error in observational studies. They do not have an error term; the only random variables with an error term in the regression equation are ϵ and Y.

Jacob: GLM stands for generalized linear model, implying that the relation is linear. But GLMs are used when the response variable is not a linear function of the explanatory variables.

Rachel: The term *generalized linear* means that the response variable is a function of a linear combination of the explanatory variables. This function is inverse of the link function. That is, the link function of the fitted response variable is a linear function of the explanatory variables.

Part B: The variance of the error term is the same at each X value: var (ϵ_j) = σ^2_ε = constant for all x_j.

Illustration: If σ^2_ε = 2 at x = 1, then σ^2_ε = 2 at x = 3

Jacob: Another textbook says that the variance of the error term is the same at each expected Y value.

Rachel: E(Y) = $\alpha + \beta \times X$, so the expected Y values map one to one into the X values. The variance of the error term is the same for all observations.

Intuition: Generalized linear models differ from classical regression analysis several ways, including the relation of the variance of the response variable to its mean.

- For generalized linear models, the variance of y_j is often a function of $\hat{y}_j$.
- For classical regression analysis, the variance of y_j is independent of $\hat{y}_j$.

Jacob: Is constant variance a reasonable assumption?

Rachel: The assumption of constant variance may not reflect actual conditions, but it underlies the formulas for ordinary least squares estimators.

Illustration: Suppose we regress annual income of actuaries on the number of exams they have passed, with $\alpha = 30{,}000$ and $\beta = 10{,}000$.

- Students with no exams have average annual incomes of 30,000.
- Average income increases $10,000 an exam, so actuaries with nine exams have average incomes of 120,000.

This assumed relation is the population regression parameters: the relation of income to exams is linear.

Actual salaries are not the same for all actuaries at a given exam level.

- Students with no exams get about the same starting salary. The actual range is narrow, perhaps 28,000 to 32,000.
- Experienced actuaries vary: some are gifted statisticians or managers and earn 160,000, and some are poor workers and earn 80,000.

Even if the response variable is a linear function of the explanatory variable, the variance is rarely constant.

Jacob: This illustration shows that omitted variables may be more important for actuaries with more exams. It does not relate the variance to the size of the explanatory variable.

Rachel: GLMs show the relation of the variance of the response variable to its mean (fitted) value.

The relation of the variance to the mean depends on the conditional distribution of the response variable.

- General insurance claim counts have Poisson distributions or negative binomial distributions: the variance is proportional to the mean.
- General insurance claim severities have lognormal or Gamma distributions: the standard deviation is proportional to the mean.
- Life insurance mortality rates have binomial distributions: the variance is proportional to $\pi \times (1 - \pi)$.

Jacob: If the variance of the response variable depends on its mean, why not use that as our assumption?

Rachel: In real statistical applications, we often do. We use weighted least squares estimators or GLMs (which use iterated weighted least squares estimators). The assumption of constant variance underlies the formulas for ordinary least squares estimators.

Part C: Normality: The error terms have a normal distribution with a mean of zero and a variance of σ^2_ε:

$$\epsilon_j \sim N(0, \sigma^2_\varepsilon)$$

Jacob: Why not state these first three assumptions in the reverse order?

- First, the error terms have a normal distribution. This seems most important.
  - If the error terms have a Poisson distribution, Gamma distribution, or binomial distribution, their variance depends on the expected value of the response variable.
  - If the error terms have a normal distribution, their variance may or may not be constant.
- Second, the variance of the error term is the same at all points. This assumption underlies the formulas for the ordinary least squares estimators.
  - If the error terms have a normal distribution with constant variance, the least squares estimators are also maximum likelihood estimators.
- Third, the mean of the error term is zero. If it equals $k \neq 0$, add k to α to get a mean error term of zero.

Rachel: On the contrary: the first three assumptions are listed in order of importance.

- The first assumption is critical if the expected error term is not zero at each point, the response variable is not a linear function of the explanatory variables. The model is not correctly specified, so it is biased.

Jacob: Is the model biased only if it is structural, not if it is empirical?

Rachel: Fox says: *Omission of explanatory variables* causes bias only for structural relations, not for empirical relations. *Model specifiication error* causes bias for both types of relation.

- The second assumption is also important: if the variance of the error term is not constant, we should give greater weight to points with lower variance when fitting the regression line. The simple functional forms for A and B (the ordinary least squares estimators for α and β) would not be correct.

- The third assumption is not that important. If the mean of the error term is zero and its variance is constant, the regression line is still the most efficient estimator along all unbiased linear estimators even if the distribution of the error terms is not normal.

Jacob: If the assumption about a normal distribution is not important, why do we include it?

Rachel: This assumption is underlies the formulas for the *standard errors* of the regression coefficients, the *t*-values, the *p*-values, and the confidence intervals.

- If the error terms have a normal distribution with constant variance, we exact standard errors, *t*-values, *p*-values, and confidence intervals.
- If the error terms do not have a normal distribution, these values are not exact. They are asymptotically correct for large samples (under reasonable conditions), but they are not exact.
- The standard errors, *t*-values, *p*-values, and confidence intervals are not exact, but they are reasonably good *if the variance of the error terms is constant*. If this variance is not constant, we use weighted least squares or generalized linear models.

Jacob: Is this assumption of normally distributed error terms generally true?

Rachel: The central limit theorem implies that it is a good approximation if the explanatory variables are the sum of independent random variables of similar size.

Jacob: Another textbook says that the Y values have a conditional normal distribution.

Rachel: $y_j = \alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk} + \epsilon_j$. The only random variable on the right side of the equation is ϵ_j, so if ϵ_j has a normal distribution, y_j has a normal distribution. Fox says: “Equivalently, the conditional distribution of the response variable is normal: $Y_j \sim N(\alpha + \beta x_j, \sigma^2_\varepsilon)$.”

Jacob: How does a conditional distribution differ from a distribution?

Rachel: The term *distribution* (or *probability distribution*) has two meanings. The regression analysis formulas use both meanings, which is confusing to some candidates.

The sample distribution is the distribution of the observed values in the sample. The values of $\sigma^2(x)$ and $\sigma^2(y)$ are the sample distributions of the observed x- and y-values.

The conditional distribution is the distribution of the random variable about its expected value. The explanatory variable is chosen by the statistician (in experimental studies) or observed without error in observational studies, so it has no conditional distribution. The response variable is a random variable. It has a conditional distribution at each point.

Jacob: How do these distributions differ for the error term?

Rachel: The error terms are independent and identically distributed. Each error term is normally distributed with a mean of zero and a variance of σ^2_ε, so the distribution of all error terms is also normally distributed with a mean of zero and a variance of σ^2_ε.

Take heed: This is the distribution of the population of error terms. The sample distribution of the observed values does not have a mean of zero and a variance of σ^2_ε (unless $N \to \infty$). The sample distribution of the residuals has a mean of zero (by construction) and a variance whose expected value is σ^2_ε but whose actual value may differ.

Jacob: How do these distributions differ for the response variable?

Rachel: The expected value of y_j is $\alpha + \beta \times x_j$. The sample distribution of the y-values does not have a normal distribution. Each y_j is $\alpha + \beta \times x_j + \epsilon_j$, so each response variable has a normal distribution with a mean of its expected value and a variance of σ^2_ε.

Independence of error terms vs time series

Part D: Independence: the error terms are independent: ϵ_j, ϵ_k are independent for $j \neq k$

Jacob: Independence seems generally true. If we regress home prices on home sizes, why would the error term for observation *j+1* depend on error term for observation *j*?

Rachel: We have two statistics on-line courses: one for regression analysis and one for time series. In a time series, random fluctuations persist at least one more period. The formulas for confidence intervals of forecasts differ from the formulas in the regression analysis course.

Jacob: Why does a regression where the response variable is a time series create problems?

Rachel: If the response variable is a time series, the expected value of the error term is not zero at each point, and the error term at point *j+1* depends on the error term at point *j*. Suppose we regress the daily temperature on explanatory variables like cloud cover, amount of smog, carbon dioxide emissions, and humidity. These explanatory variables may affect the daily temperature, but we have omitted the effect of the previous day's temperature. The least squares estimate for the variance of the error term is too large.

Jacob: A time series looks at values over time. What if all data points are from the same time, such as daily temperatures from the same day at different places?

Rachel: A time series can be temporal or spacial. The daily temperature is similar in adjacent cities and less similar at the distance between the cities increases.

Illustration: Home prices may be a linear function of plot size and the number of rooms. But home prices have both spacial and temporal relations as well.

- Homes in adjacent locations have more similar prices than homes in different states or countries.
- Home prices for all properties fluctuate over time with interest rates, recessions, and tax rates.

Illustration: Suppose we measure daily temperature in a city on the equator. The expected daily temperature is 80° every day of the year, with on seasonal fluctuations, but random fluctuations cause the temperature to vary from 60° to 100°. Daily temperature is a time series:.if the temperature is high one day, it will probably be high the next day as well, since the high temperature reflects weather patterns that persist for several days.

The Fox textbook says: “The assumption of independence needs to be justified by the procedures of data collection. For example, if the data constitute a simple random sample drawn from a large population, then the assumption of independence will be met to a close approximation. In contrast, if the data comprise a time series, then the assumption of independence may be very wrong.”

Jacob: Another textbook says that any two Y values are independent: each y_j is *normally distributed and independent.* Does this mean

- that Y_j - Y_k is independent of j - k (that is, Y_{j+1} is independent of Y_j), or
- that Y_j - $\hat{y}_j$ is independent of Y_k - $\hat{y}_k$?

Rachel: $y_j = \alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk} + \epsilon_j$. The only random variable on the right side of the equation is ϵ_j, and $\hat{y}_j = \alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk}$, so if ϵ_j and ϵ_k are independent, y_j - $\hat{y}_j$ and y_k - $\hat{y}_k$ are independent.

The Fox textbook says: "Any pair of errors ϵ_i and ϵ_j (or, equivalently, conditional response variables Y_i and Y_j) are independent for $i \neq j$.

Explanatory variables: fixed or measured without error

Part E: The explanatory variables (X values) are fixed.

- In experimental studies, the statistician picks explanatory variables and observes the response variable.
- In observational studies, the explanatory variables are measured without error.

Jacob: Are actuarial pricing studies experimental studies?

- To see how sex affects claim frequency, one picks 100 male drivers and 100 female drivers.
- To see how age affects claim frequency, one picks 100 youthful drivers and 100 adult drivers.

Rachel: Actuarial pricing studies are observational studies, not experimental studies. The actuary observes male drivers and female drivers; the actual does not create male drivers and female drivers. Sex and age are attributes of the driver, not interventions.

Jacob: What is an intervention?

Rachel: In experimental studies, the researcher creates the explanatory variable. A clinical trial for a new drug uses 100 randomly selected subjects show are given the drug and 100 randomly selected subjects are are not given the drug (or given a placebo). Any subject may be shifted between the two groups (control group vs treatment group). In contrast, a driver can not be shifted between male and female drivers by changing sex.

Jacob: Why is the difference between attributes and interventions important?

Rachel: Attributes are often correlated with other (omitted) explanatory variables. Attributes may create biases in structural relations; interventions are less likely to create biases if they are truly random.

Illustration: An actuary regresses motor insurance claim frequency on urban vs rural territories and concludes that urban drivers have higher claim frequency than rural drivers. But territory is not the cause, so the study is biased. The actual causes of the territorial difference are the attributes of residents of cities vs rural areas.

Illustration: Duncan's Canadian occupational prestige is an observational study, as are most social science and actuarial studies.

Observational studies

In observational studies, the explanatory variables values are observed, not fixed by design.

Jacob: Are the observed x-values a random sample?

Rachel: The term random sample has several meanings. An actuary comparing urban vs rural drivers may use all drivers in the insurer's data base or a random sample of those drivers. But this sample (or population) is not random. The insurer may write business in states or countries where urban drives are poorer than rural or suburban drivers. The regression analysis may not apply to places where urban drivers are richer than rural or suburban drivers. In some European and Asian countries, urban residents are wealthier than suburban or rural residents.

The statistical term *randomized* means that an intervention is applied randomly to subjects. The subjects have no other differences than the random application of the intervention.

Illustration: Patients are randomly assigned the new vs the old medication. Even the doctors and nurses do not know which patients receive which medication.

Accurate measurement

In observational studies, explanatory variables are measured without error and are independent of the errors. If the explanatory variables are random, or if we do not measure them precisely, our estimates have another source of random fluctuation.

Illustration: A statistician regresses home values on area of the house or the lot. The regression depends on accurate measurement of the area. We assume the area is measured without errors.

Jacob: Is this assumption reasonable?

Rachel: It is often reasonable. Life insurance mortality rates and personal auto claim frequencies depend on the age and sex of the policyholder (or driver). Age and sex are not random variables, and we measure them without random errors. Age and sex are attributes, not interventions, so they are not independent of the causal variables affecting mortality rates and claim frequency, but they are measured without error,.

Part F: X is not invariant: the data points have at least two X values. If the X value is the same for all data points, the least squares estimators are $\alpha = \overline{y}$ (the mean of the y-values).

Jacob: Do we infer that $\beta = 0$ if all the x-values are the same?

Rachel: We can not infer anything about β, since no data show how a change in x affects y.

Jacob: Does invariance that mean that no X values should be the same?

Rachel: On the contrary, many X values are the same in regression analyses. Suppose we regress personal auto claim frequency on sex and age of the driver and the territory in which the car is garaged. Sex has two values, age may be one of three values (youthful, adult, retired), and the state may have ten territories. If the regression uses 100,000 cars, many sets of X values are identical.

** Exercise 11.2: Expected values

Which of the following statements stem from the assumptions of classical regression analysis?

x_j is the value of the explanatory variable at point j.
ϵ_j is the value of the error term at point j.
y_j is the observed value of the explanatory variable at point j.
$\hat{y}_j$ is the fitted value of the explanatory variable at point j.

A. The correlation of x_j with ϵ_j is zero.
B. The correlation of $\hat{y}_j$ with ϵ_j is zero.
C. The correlation of y_j with ϵ_j is zero.

Part A: Classical regression analysis assumes the error term is independent of the explanatory variables.

Part B: $\hat{y} = \alpha + \beta \times x$, so the correlation of $\hat{y}$ with ϵ is $\rho(\epsilon, \alpha + \beta \times x) = \rho(\epsilon, \alpha) + \beta \times \rho\ (\epsilon, x) = 0 + \beta \times 0 = 0$.

Part C: $y = \hat{y} + \epsilon$, so $\rho(y, \epsilon) \neq 0$. When $\epsilon > 0$, $y > \hat{y}$, and when $\epsilon < 0$, $y < \hat{y}$. We state two implications:

- For any given observation *j*, ϵ_j and y_j are random variables that are perfectly correlated.
- For all observations combined, ϵ_j and y_j are random variables that are positively correlated.

** Exercise 11.3: Expected values

Which of the following statements stem from the assumptions of classical regression analysis?

x_j is the value of the explanatory variable at point j.
ϵ_j is the value of the error term at point j.
y_j is the observed value of the explanatory variable at point j.
$\hat{y}_j$ is the fitted value of the explanatory variable at point j.

A. The expected value of $x_j \times \epsilon_j$ is zero.
B. The expected value of $y_j \times \epsilon_j$ is zero.
C. The expected value of $\hat{y}_j \times \epsilon_j$ is zero.

Solution 11.3: The solution to this exercise follows from the previous exercise.

Part A: $\rho(x_j, \epsilon_j) = 0$, so $E(x_j \times \epsilon_j) = 0$

Part B: $\rho(y_j, \epsilon_j) > 0$, so $E(y_j \times \epsilon_j) \neq 0$

Part C: $\rho(\hat{y}_j, \epsilon_j) = 0$, so $E(\hat{y}_j \times \epsilon_j) = 0$

** Exercise 11.4: Least-squares coefficients

A regression model with N observations and k explanatory variables is

$$y_j = \alpha + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots + \beta_k x_{jk} + \epsilon_j.$$

Under the assumptions of classical regression analysis, the least-squares coefficients have five attributes. Explain each of these attributes.

A. Linear functions of the data
B. Unbiased estimators of the population regression coefficients
C. The most efficient unbiased linear estimators of the population regression coefficients
D. The same as the maximum likelihood estimators
E. Normally distributed

Part A: The least squares coefficients are linear functions of the observed data.

Jacob: How are they linear functions? $B = \sum(x_i - \overline{x})(y_i - \overline{y}) / \sum(x_i - \overline{x})^2$. This has terms of x_i y_i in the numerator and x^{2i} in the denominator. $A = \overline{y} - B \times \overline{x}$, so A is also a function of both the x and y values.

Rachel: The *observed data* are the response variable (the Y values), not the explanatory variables (X values).

- The explanatory variable is fixed, not a random variable.
- The error term is a random variable is a fixed (but usually unknown) variance σ^2_ε.
- α and β are unknown population regression parameters, not random variables which we estimate.
- The response variable y_j is a random variable with
  - a mean of $\alpha + \beta \times x_j$
  - a variance of σ^2_ε

The least squares estimator B is a linear function of the y_j.

- If we repeat the regression analysis, B will differ from the first analysis, since the y-values differ.
- The x-values do not differ in the two experimental studies.
- If we repeat the regression analysis many times, the mean B values approaches β.

The x_j are the coefficients of the linear function of the y-values.

Jacob: Must we know these coefficients for the on-line course?

Rachel: The fitted values $\hat{y}_j$ are $A + B\, x_j$. A and B are linear functions of the observed y-values, so each $\hat{y}$ value is also a linear function of all the observed y-values. The module on hat-values gives the coefficients of this linear function.

Jacob: Most regression analysis in the social sciences (and in actuarial work) are observational studies, not experimental studies, where the x-values are chosen from observed values. Aren't they random variables?

Rachel: Even for observational studies, where X values are observed (not fixed by the statistician), the values are measured without error. They are not random variables, since they have no standard error.

Illustration: An actuary regresses life insurance mortality rates on sex and age. The observed mortality rate is a random variable (with a binomial distribution). Sex and age may be sampled from the insurer's data, but they are fixed quantities with no measurement error.

Illustration: A real estate broker regresses home prices on the number of rooms and the size of the home. The home price is a random variable, which may fluctuate from day to day, depending on the offers by potential buyers. The number of rooms and the size of the home are depend on the homes available for sale but they are fixed quantities with no measurement error.

Jacob: What difference does measurement error make?

Rachel: Classical regression analysis assumes $Y_j = \alpha + \beta X_j + \epsilon_j$. If the X_j have measurement error, we are regressing Y_j on (X_j + the measurement error). The inferences of the regression analysis, such as the variance of the error terms, no longer hold.

Part B: The least squares estimator are unbiased estimators of the population regression coefficients.

Jacob: β is a fixed but unknown population regression parameter; B is a random variable.

- If B = β, then B is correct.
- If B > β, B is too high.
- If B < β, B is too low.

If B > β, doesn't that mean that B is biased upward? If B < β, doesn't that mean that B is biased downward? We want the correct figures, not figures that are too high or too low.

Rachel: B is a random variable, with errors that are normally distributed. B is never exactly equal to β because of random fluctuations, but the expected value of B is β.

Don't confuse an **over-estimate** or **under-estimate** with a biased **estimator**.

- An *estimate* may be too high or too low even if it is unbiased.
- An *estimator* can be biased or unbiased.

Illustration: Suppose the true linear relation is

motor insurance accident frequency = 0.0005% × distance driven each year (in miles or kilometers).

A regression gives B = 0.0007% on one sample and 0.0004% on another sample. If we repeat the regression an unlimited number of times, the average B should be 0.0005% (if the assumptions of classical regression analysis hold).

Jacob: Aren't most estimators unbiased? Why would anyone use a biased estimator?

Rachel: Some estimators are unbiased, some are biased. Actuaries use biased estimators for many studies, since they may reduce the mean squared error.

Illustration: An actuary examines claim sizes, where the size-of-loss distribution is lognormal. Some actuaries use the median or a mean excluding the highest and lowest values to avoid distortion by random large losses. These estimators are useful, especially if the sample is small, but they are biased. The median and the ex-high-low mean understate the true mean of lognormal distribution.

Part C: Suppose we have a random sample of ten numbers from a normal distribution with a mean of μ. We estimate the mean of the normal distribution three ways:

- Method #1: The average of the highest and lowest numbers.
- Method #2: The average of all numbers except the highest and lowest numbers.
- Method #3: The average of all the numbers.

If the numbers are normal distributed, all three methods are unbiased. We repeat the random sample 10,000 times, getting three estimators for each sample. Each method gives 10,000 estimates.

Let ${}_{k}B_{j}$ be estimate #j for method #k. For example, ${}_{2}B_{1,000}$ is the 1,000th estimate for method #2. If all methods are unbiased, $\sum {}_{1}B_{j}$ / 10,000 = $\sum {}_{2}B_{j}$ / 10,000 = $\sum {}_{3}B_{j}$ / 10,000 = μ.

The squared error is $({}_{k}B_{j} - \mu)^{2}$, and the mean squared error is $\sum ({}_{k}B_{j} - \mu)^{2}$ / 10,000. The three estimators do not have the same mean squared error. Rather

$$\sum({}_{1}B_{j} - \mu)^{2} / 10{,}000 > \sum({}_{2}B_{j} - \mu)^{2} / 10{,}000 > \sum({}_{3}B_{j} - \mu)^{2} / 10{,}000.$$

An estimator with a lower mean squared error is more efficient. The least squares estimators are the *most* efficient *linear* estimators.

Jacob: Perhaps one of the samples has an unusually high figure that occurs rarely. In this case, Method #2 might have a lower mean squared error than Method #3.

Rachel: You are right. Instead of 10,000, we should say N estimates. In the limit, as N → ∞,

$$\sum({}_{1}B_{j} - \mu)^{2} / N > \sum({}_{2}B_{j} - \mu)^{2} / N > \sum({}_{3}B_{j} - \mu)^{2} / N.$$

Jacob: Why is the word *linear* in the line: the least squares estimators are the *most* efficient *linear* estimators?

Rachel: Other estimators, such as maximum likelihood estimators, may be more efficient, but they are not linear functions of the observed values. (But see the answer to Part D.)

Part D: The maximum likelihood estimators are the best estimators. Maximum likelihood is covered in later modules, not here. *If the classical regression assumptions are true*, the least squares estimators are also the maximum likelihood estimators.

Jacob: Part C says that non-linear estimators (like maximum likelihood estimators) may be more efficient than ordinary least squares estimators. Part D says that the maximum likelihood estimators are the best, and that ordinary least squares estimators are the same as the maximum likelihood estimators. Is this contradictory?

Rachel: Part C assumes linearity, constant variance, and independence; it does not assume that the error terms have a normal distribution. Part D assumes the error terms have a normal distribution. Fox says: *when the error distribution is heavier tailed than normal, the least squares estimators may be much less efficient than certain robust-regression estimators, which are not linear functions of the data.*

Part E: The least squares estimators are normally distributed.

Jacob: An estimate is a scalar, like B = 1.183 or B = 25,000. B doesn't have a distribution.

Rachel: Suppose the true β is 100. We take 10,000 samples of 10 data points each, and for each sample we compute B. The 10,000 B's have a normal distribution with a mean of β.

Jacob: Why do we care about the distribution of the ordinary least squares estimators?

Rachel: The standard errors of the estimators, *t*-values, *p*-values, and confidence intervals assume a normal distribution for the estimators. If the true distribution is not normal, the estimator may still be unbiased, but the formulas to calculate standard errors of the estimators, *t*-values, *p*-values, and confidence intervals are not correct.

** Exercise 11.5: *Empirical association* vs *causal relation*

Fox distinguishes between *empirical association* and *causal relation*.

If an explanatory variable X_2 is omitted from a regression equation, the incomplete regression equation (using only X_1) is biased if the following three conditions are true. Explain each of them.

A. The regression equation represents a causal relation (a structural model), not an empirical association.
B. The omitted explanatory variable X_2 is a cause of the response variable Y.
C. The omitted explanatory variable X_2 is correlated with the explanatory variable X_1.

Part A: Suppose the response variable is the annual trend in workers' compensation loss costs, and the two explanatory variables are X_1 = wage inflation and X_2 = medical inflation. Workers' compensation has two pieces: indemnity benefits, which increase with wage inflation, and medical benefits, which increase with medical inflation. For simplicity, suppose indemnity and medical benefits are each 50% of total benefits, and no other items besides wage and medical inflation affect workers' compensation loss costs.

The proper regression equation is $Y = 0.500\ X_1 + 0.500\ X_2 + \epsilon_j$

Wage inflation and medical inflation are correlated, since both reflect monetary inflation. Suppose their average values are 5% per annum and their correlation is 80%.

An actuary uses the regression equation $Y = \alpha + \beta_1\ X_1 + \epsilon_j$

The least squares estimators are A = 1% and $B_1 = 0.900$.

Question: Is this regression equation biased?

Answer: If the regression equation represents an empirical association, it is correct. Wage inflation (X_1) has a 90% correlation with workers' compensation loss costs: 50% through indemnity benefits and 50% × 80% = 40% through medical benefits. If the regression equation represents a causal relation, it is biased. If wage inflation increases 1% but medical inflation does not change, workers' compensation loss costs increase 0.5%, not 0.9%.

Jacob: The population regression parameter α is zero, so shouldn't A be zero as well?

Rachel: Only 80% of medical inflation affects the regression equation with a β_1 parameter but no β_2 parameter. The average medical inflation is 5% per annum, so 20% × 5% = 1% is included in A.

Part B: The second and third conditions are that

- The omitted explanatory variable X_2 is a cause of the response variable Y.
- The omitted explanatory variable X_2 is correlated with another explanatory variable X_1.

Question: Why does this say that X_2 is a *cause* of the response variable Y but is just correlated with X_1?

Answer: Suppose we regress workers' compensation loss cost trends on inflation. Assume (for simplicity) that inflation causes the loss cost trend and it also determines the nominal interest rate. That is, 1% higher inflation causes a 1% higher trend and 1% higher nominal interest rate. In this example, X_1 = inflation, X_2 = interest rate, and Y = loss cost trend. X_2 is omitted from the regression equation, and it is correlated with Y and X_1, but no bias exists, since interest rates do not cause loss cost trends.

** Exercise 11.6: Bias

A statistician regresses Y on explanatory variable X_1 but does not use a second explanatory variable X_2. The regression line is $Y = \alpha + \beta X_1 + \epsilon$

Under what combination of the following conditions is the estimate of β biased?

A. $\rho(Y, X_1) = 0$
B. $\rho(Y, X_2) = 0$
C. $\rho(Y, X_1) \neq 0$
D. $\rho(Y, X_2) \neq 0$
E. $\rho(X_1, X_2) = 0$
F. $\rho(X_1, X_2) \neq 0$
G. X_1 has a causal effect on Y
H. X_2 has a causal effect on Y
I. X_1 does not have a causal effect on Y
J. X_2 does not have a causal effect on Y
K. The regression equation is structural (the explanatory variables cause the response variable)
L. The regression equation is associative (explanatory variables are correlated with the response variable)

Solution 11.6: F, H, and K

Suppose higher money supply growth causes higher inflation, which causes higher nominal interest rates.

Type of relation: An associative relation gives the empirical relation of X_1 and Y. Ignoring other variables is not relevant. Nominal interest rates are empirically related to the money supply growth rate, so a regression on interest rates on money supply growth is fine. But inflation is the cause of higher nominal interest rates. The regression of nominal interest rates on the money supply growth rate is not a proper causal relation. It is biased, since it ignores the effects on inflation.

Causal effect: The nominal interest rate and the inflation rate are correlated. A structural regression of interest rates on the money supply growth rate is biased, since the omitted variable (inflation) is a cause of higher interest rates. A structural regression of inflation rates on the money supply growth rate is not biased, since the omitted variable (interest rates) is not the cause of higher inflation rates.

Correlation: A structural regression of interest rates on the money supply growth rate is biased, since the omitted variable (inflation) is correlated with the money supply growth rate.

** Question 11.7: Observational studies

In observational studies, the X values are sampled, not fixed by design. In these studies, which of the following is an assumption of classical regression analysis?

A. The explanatory variable is measured without error and is independent of the error.
B. The explanatory variable and the error are independent and unbiased.
C. The explanatory variable and the error are independent and efficient.
D. The error is a linear function of the explanatory variable.
E. The regression assumptions are not satisfied in observational studies.

Answer 11.7: A

Jacob: What are observational studies vs experimental studies?

Rachel: Suppose we want to assess the effects of diet and exercise on weight gain or loss.

In an experimental study, we specify the diet and exercise regime for N persons. The first person might be given a diet of 1,000 calories each of carbohydrates, fat, and proteins, with a 30 minute exercise regime. The second might be given a diet of 1,500 calories each of carbohydrates and proteins (with no fats), with a 60 minute exercise regime. For each person, the explanatory variables are fixed by the research study. We then observe the weight gain or loss during the period.

In an observational study, we observe N persons. For each person, we record the diet and the exercise each day. We then observe the weight gain or loss during the period.

Jacob: What is the advantage of observational studies?

Rachel: It is relatively easy to record the diet and exercise of each person. Each person records all food and exercise in a diary. It is not easy to specify what each person will eat every day, or what exercise each person will do.

Jacob: What are the drawbacks of observational studies?

Rachel: The drawbacks are measurement error, spread of the explanatory variables, and bias.

Measurement error: People tend to mis-estimate or ignore personal data. In a study of weight gain, a person may say he had a small slice of pizza instead of a large one or may ignore a ice cream sundae.

Spread: Most people have similar percentages of carbohydrates, fats, and protein. Observational studies with low variances of the explanatory variables have high standard errors of the regression coefficients.

Bias: People who eat more are often larger, so they may gain more weight. To hold constant other explanatory variables, experimental studies randomly assign diets to a sample of people.

Jacob: For experimental studies, are the X values random variables?

Rachel: For experimental studies, the X values are fixed by design.