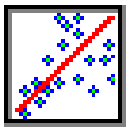
Solutions to Worktext Exercises



Chapter 16

Visualizing Regression Assumptions

Basic Learning Exercises

1. The scatter plot is scattered around the estimated regression line in a "fan-out" fashion. This same pattern is seen more clearly on the bar graph. The "fan-out" pattern indicates heteroskedasticity because the disturbances on the left side have less variability about the regression line than those on the right side.

2. Both the scatter plot and bar graph show large residuals on the left side of the graph and small residuals on the right side. The "funnel-in" pattern indicates heteroskedasticity because the disturbances on the left side of the graph are spread further away from the regression line than those on the right side.

3. a) The variance of the error term increases with X. b) With low money growth, we are unlikely to have high inflation, whereas with high money growth, although you will probably have high inflation, inflation could be low. c) An advantage of a scatter plot is it would reflect the value (not the order) of X, providing additional information. However, a bar more clearly displays the residual's deviation from the fitted line. d) The variance of the error term decreases as Y decreases. e) Since the heteroskedasticity is related to E(Y) which is only a function of X, the relationship is clearly visible in this case, however, if there were more than one X variable it probably wouldn't be.

4. The residuals appear to be spread randomly about the regression line and on the bar graph. The heteroskedasticity is not revealed because it is not related to the X variable. It would only show up if the residuals were plotted against the 3rd variable causing the heteroskedasticity.

5. a) The scatter plot shows runs where the data points are above or below the estimated regression line. This is more evident on the bar graph where there are streaks of positive and negative residuals. b) 90% or more of the time on the bar chart, slightly less on the scatter plot. c) This is called first-order autocorrelation because each disturbance term is a function of the one that proceeds it. d) The order of the observations is meaningful in time series data. In cross sectional the data is often alphabetical (state or country data) or random (sample of households). e) Even in time series data, the X values don't always increase. If this happens, the order is changed when the ordered X values are used rather than the observation number.

6. a) The scatter plot shows runs where the data points are above or below the estimated regression line. This is more evident on the bar graph where there are runs of positive and negative residuals. It is not as clear as in question 5. b) About 70% of the time it is clear and about 20% of the time it is suspected on the bar chart. It is less evident on the scatter plot. $\rho = 0.5$. c) About 40% of the time it is clear and about 40% of the time it is suspected on the bar chart. It is less evident on the scatter plot. d) As rho decreases, autocorrelation is more difficult to detect visually.

7. a) The residuals alternate between positive (above fitted line) and negative (below fitted line). It is clearer on the bar graph. b) When $\rho \geq -0.25$ it is difficult to detect visually. c) The real world changes slowly, building on what proceeded it, not reacting to it.

8. a) The histogram has the usual bell-shape. Standardized residuals rarely lie in the first or last interval. It is common for there to be values in the interval between –3 and –2 or in the interval between 2 and 3. b) The standardized residuals consistently lie close to the 45-degree line. It is rare for any values to be below –3 or above 3.

9. a) Between –2 and –1.5 the residuals consistently have a tail below the 45-degree line, between 1.5 and 2, the tail is above the line. b) The histogram usually uses only the middle four intervals since it is unusual for standardized residuals to be below –2 or above +2. c) Yes, both could be used, but the normal probability plot is more reliable.

10. a) The histogram is very peaked (the intervals –1 to 0 and 0 to 1 contain most of the observations). It is common to have standardized residuals below –3 and above 3. b) The probability plot generally shows values above the 45-degree line below –3 and below the line above 3. c) Both could be used, but the normal probability plot is more reliable.

11. a) The standardized residuals generally lie along the line between –2.5 and 2.5. There is a tendency for the residuals to be below the line around –2.5 and above the line around 2.5. b) The histogram is approximately symmetric with no standardized residuals below –3 or above 3. c) There appear to be too few observations in the –1 to 0 and 0 to 1 intervals. d) No. Only if you knew it wasn't normally distributed could you see it.

Intermediate Learning Exercises

12. a) $J = -50 + 2\,S + u$ b) Normal with a mean of 2 and a variance of $1600 / \Sigma(x - \bar{x})^2$. c) Since the histogram is centered on it true value 2, it is unbiased. d) Since the histogram is outlined by the normal distribution, it is normally distributed. e) Since the histogram is outlined by the normal distribution, it is normally distributed. f) Since the histogram is centered on –50, it is unbiased.

13. Slope: Minimum __–5__ Maximum __9__ Intercept: Minimum __–640__ Maximum __640__ The slope's minimum is now –0.1 and its maximum is 4.1. The intercept's minimum is now –300 and its maximum is 180. Since both estimators are unbiased and since the variance of both estimators decreased when n increased, both estimators are consistent.

14. The theoretical distribution outlines the histogram hence the estimators are efficient.

15. a) The number of rejections in both tails is approximately 25 (2.5% of 1000). b) ±2 c) Clearly over half of the distribution is outside the critical values, so about 65%.

16. Because the histogram for the slope is centered on 2 and for the intercept on 0, both are unbiased. Superimposing the normal curve shows that both are normally distributed, although with an increased variance.

17. Although both estimators are unbiased (first requirement), because the histogram for the *variance* is wider than the theoretical distribution, the estimators are not efficient.

18. The t-statistics will be wider than expected since the variance is more spread out than it should. Hence, if the null is true (β_0), H_0 will be incorrectly rejected more often than the α-level (biased test). If the null is not true (β_1), H_0 will be under-rejected (less power).

19. Yes, they are both consistent since both are unbiased and each of their variances increased as the sample size was decreased. The range of the slope increased from about 4 to 12 when the sample size was reduced from 100 to 10, while the intercept's increased from about 400 to 1000.

20. No, all the conclusions are the same. The fact that they are more dramatic is a function of this particular degree of heteroskedasticity, not funneling-in.

21. Because the histogram for the slope is centered on 2 and for the intercept on 0, both are unbiased. Superimposing the normal curve shows that both are normally distributed, although with an increased variance.

22. Although both estimators are unbiased (first requirement), because the histogram for the *variance* is wider and shifted to the right of the theoretical distribution, the estimators are not efficient.

23. The t-statistics will be wider than expected since the variance is more spread out than without autocorrelation. Hence, if the null hypothesis is true (β_0), H_0 will be incorrectly rejected more often than the α-level (biased test). If the null is not true (β_1), the t-statistic will be closer to zero than without autocorrelation since the variance σ^2 is on average larger than expected causing the t-statistic to be less, in absolute value, than expected (the square root of the variance is in the denominator of the t-statistic). Therefore, H_0 will be under-rejected (less power).

24. Yes they are both consistent since both are unbiased and each of their variances increased as the sample size was decreased. The range of the slope increased from about 5 to 16 when the sample size was reduced from 100 to 10, while the intercept's range increased from about 440 to 1280.

25. The histograms of the estimated slope and intercept and their t-statistics are narrower than before. If the null hypothesis is true (β_0), it will be rejected much less often than the α-level (biased test). If the null hypothesis is false (β_1), the t-statistic is closer to zero since the variance is larger than with no autocorrelation. This reduces the power. However, the narrower histogram increases the power. The tradeoff depends on the particular model and value of rho.

26. a) Both estimators are unbiased since the histogram for the slope is centered on 2 and for the intercept on 0. b) Neither estimator is normally distributed. c) Although both estimators are unbiased (first requirement), because the histogram for the *variance* is shifted left of the theoretical distribution, the estimators are not efficient.

27. a) When the null hypothesis is true (β_0), the distribution appears to have the appropriate Student's t distribution. If the null is not true (β_1), the t-statistic is shifted right because the variance σ^2 is on average smaller than expected causing the t-statistic to be larger than expected (the square root of the variance is in the denominator of the t-statistic).
b) Therefore, H_0 will be over-rejected (more power). c) Both estimators are consistent since both are unbiased and their variances decreased as the sample size was increased. The range of the slope decreased from about 14 to 4.2 when the sample size was increased from 10 to 100, while the intercept's range decreased from about 1200 to 400.

28. Yes, it appears that both are normally distributed. When the null hypothesis is true (β_0), the test statistic seems to be distributed as Student's t. However, when the null hypothesis is false (β_1), the test statistic is larger than expected (increased power).

29. a) Both estimators are approximately normal. b) Both seem to be distributed as Student's t.
c) Regression estimators are robust against some non-normality.

Advanced Learning Exercises

30. a) $R = 5 + 1\ K + 0.2\ Y$ b) $R = \beta_0 + \beta_1 K$ c) $R = 6.76 + 0.985\ K$ or something similar. d) The true model (black line) is below the estimated line.

31. a) Both parameters are biased. b) Because both parameters are biased, neither can be efficient nor consistent. c) The model estimated is wrong and is of questionable value.

32. a) The mean of Z, years until maturity, is 8. b) Its coefficient is 0.2. c) The estimator for the slope is unbiased, normally distributed, consistent, but not efficient. d) The bias of 1.6 equals the mean of Z times its coefficient (8×0.2).

33. The true equation is $R = 5 + 1\ K + 0.2\ Y$. Since Y and K are perfectly correlated (recall that X is K and Z is Y), this can be written as $R = 5 + K\ (1 + 0.2) + 0.2\ (\bar{Y} - \bar{K})$. Hence, the bias for the slope coefficient is 0.2 and if the difference in the means of Y and K is not zero, the intercept will be biased.

34. *Case1: Omitted variable not related to included variable* — the estimated slope has all of the desirable properties except efficiency and the intercept has no desirable properties if the mean of the omitted variable is nonzero. *Case 2: Omitted variable perfectly correlated with included variable* — the estimated slope is biased by the coefficient of the omitted variable and the intercept is biased if the difference in the means of the omitted and included variable is nonzero. *Case 3: Omitted and included variable are correlated, but not perfectly* — both the intercept and slope estimators will be biased.

35. a) $U = 0.05\ N$ b) The X's are not fixed but stochastic. c) The X values change. d) The X values don't change. e) That the X values can change randomly.

36. The estimators are unbiased, normally distributed, consistent, and efficient.

37. a) $Q = 20 - 0.5\ P$ b) Independence between P and the error term. c) No problems are evident. d) The estimated line is consistently above the true model. e) No.

38. The estimated slope is biased, inconsistent, and inefficient. The estimated intercept is unbiased, normally distributed, consistent, and efficient.

39. a) $Q = 129 + 62\ Ln(E) + u$ b) $Q = \beta_0 + \beta_1 E + u$ c) Nothing is obviously wrong, but the histogram rarely has residuals outside of two standard deviations, and the bar graph often has negative residuals at both ends. d) The true model is a curved line with decreasing slope. The estimated line is usually above the true model at both ends and below the true model in the middle. e) Yes, since the function is different.

40. a) The true model has an intercept much lower than the estimated model and becomes increasingly different than the estimated model outside of the sample range. b) Using the wrong model yields very inaccurate forecasts. c) A 1-percent change in E (employment) causes Q (output in thousands of units) to change by 0.62 (= 62/100) or 620 units. d) If the estimated slope is 0.035, then a 1-unit change in E causes Q to change by 0.035 or 35 units.

41. a) The parameters are extremely biased. b) On average, the bias is about 400 for the intercept and about 60 for the slope. c) The estimators cannot be efficient or consistent since they are biased.

42. When the wrong functional form is used, the confidence intervals are no longer valid. In addition, it is clear that on average your estimated model can give misleading information about the underlying model.

43. a) The histogram for the estimated variance and t-statistic for the slope are similar to the true distribution (if the linear-log model was estimated), but that the t-statistic for the intercept is quite different. b) The t-statistic is the ratio of the estimated slope (which is a function of X) divided by the square root of $\sigma^2 \Sigma(X - \overline{X})^2$, where X represents Ln(E). Therefore, the difference due to the transformation is cancelled out in the numerator and denominator, and, since the variance is almost the same, the t-statistic for the slope is about correct. c) Although the wrong model is used, the linear model's estimated line is similar to the true model's estimated line. Hence, the estimate of the variance is similar.

44. The F-statistic with one independent variable is the square of the t-statistic for the slope. Since the t-statistic is about the same, the F-statistic will also be about the same.

45. The mean is about 0.65, the minimum values is about 0.28, and the maximum value is about 0.88.

46. The estimators are unbiased since the histogram is centered on the true distribution. They are each normally distributed since the histogram is outlined by the true normal distribution for each parameter.

47. They are efficient since the histogram for the variance is outlined by the true distribution for the variance.

48. They are both consistent because they are each unbiased and their variance is reduced it the sample size is increased. This is seen by a decrease in the range of each estimator's histogram.

49. Estimating a nonlinear model is no different than estimating a linear model with regard to the properties of the estimators.

50. The mean is now about 0.7, the minimum value is about 0.4, and the maximum is about 0.9. These are very similar to those in question 45. Since the Total Sum of Squares is the same in the linear model as the linear-log model (Y is the same) and since the estimated variances are similar (calculated from the Error Sum of Squares), $R^2 = 1 - ESS/TSS$ will be about the same for both models.