Solutions to Worktext Exercises

Chapter 15

Visualizing Simple Regression

Basic Learning Exercises

1. $\widehat{Wage}$ = 6.54 + 1.09 Complexity. $\hat{\beta}_0$ should be between 2 and 18, while $\hat{\beta}_1$ should be between 0.75 and 1.25.

2. If job complexity is zero, the wage is $6.54 thousand (intercept is 6.54). For each additional unit of job complexity the wage increases by $1.09 thousand (slope is 1.09).

3. Seeing the entire axis makes it easier to interpret the estimated model, especially the intercept. However, your data points take up only a small portion of the graph.

4. a) Residuals are the vertical differences between the data points and the estimated line. b) They are calculated by subtracting the observed y value from the estimated y value or ($\hat{u}_i = y_i - \hat{y}_i$). c) There is one residual for each observation or 40 in this case. d) The Ordinary Least Squares procedure guarantees that the residuals sum to zero.

5. a) By standardizing the residuals, the histogram shows the frequency of residuals within 1, 2, or 3 standard deviations of their mean (zero). b) The mean of the residuals is zero, so each residual is simply divided by the standard error of the model. c) A histogram displays the frequency of residuals within intervals, while the bar graph displays each residual or standardized residual.

6. The scatter plot that displays the residuals and the bar graph of residuals show the same thing. The bar graph has removed the trend line and the residuals are standardized for ease of labeling the vertical axis.

7. R^2 = 0.55 to 0.88 Standard Error of Model 3.5 to 6.4
β_0: Standard Error 2.3 to 4.3 t-value 0.1 to 4.5 Decision Different from zero
β_1: Standard Error 0.07 to 0.13 t-value 6 to 13 Decision Different from zero

8. The standard error of the model measures the variability of the data about the conditional mean. R^2 measures the proportion of total variability in the y values that the estimated model explains.

9. a) The parameters are β_0 and β_1. b) β_1 is the impact of a one-unit change in job complexity on a wage. c) Parameters are not observed. If we knew them, we would not need to infer them from the sample data. d) The u is the disturbance term or error that represents the effects of all factors that have not been explained by the model.

10. $\beta_0 = 10$ and $\beta_1 = 1$. The true model is Wage = 10 + 1 Complexity + u.

11. The expected value of the true model is the conditional mean of the distribution of wage given job complexity. There is error around this conditional mean.

12. Disturbances are the vertical difference between the data point and the conditional mean line. They cannot be calculated because the conditional mean is unknown.

13. The scatter plot showing the disturbances and the bar graph of disturbances are similar, except that the bar graph has rotated the trend line and standardized the disturbances for ease of labeling the vertical axis. The residuals and disturbances are similar.

14. The two histograms are almost identical. Yes, in general the residuals and disturbances have the expected distribution. However, there are differences from sample to sample.

15. R^2 will be between 0.55 and 0.88. The estimated model is a very good predictor of the true model and the residuals are a very good predictor of the disturbances.

16. R^2 will be between 0.0 and 0.40. The estimated model is still a very good predictor of the true model and residuals are very good predictors of the disturbances. Just as $\overline{Y}$is a very good estimator for μ_y even if the variance is large, the estimated model is a very good estimator for the conditional mean even if the fit is mediocre.

17. The confidence interval is created by adding and subtracting an interval from the estimated mean.

18. The confidence interval contains the true model about 90% of the time.

19. The confidence interval contains the true model about 95% of the time.

20. For a 95% confidence interval, on average, 95 out of 100 such intervals will contain the conditional mean.

21. The prediction interval for y|x lies outside the confidence interval for E(y|x).

22. Results will vary from about 3% to 7% outside the interval.

23. A 95% prediction interval will, on average, contain 95% of the observations.

Intermediate Learning Exercises

24. $J_i = -50 + 2\ S_i + u_i$, where J is job performance and S is the applicant's test score.

25. The shaded area appears to have the shape of an hourglass. The conditional mean goes through the middle of the shaded area. The interval about the conditional mean appears as an hourglass. The interval and shaded area are similar.

26. If the estimated slope is close to the true slope, the intercept is usually close to the true intercept. However, if the estimated slope is much steeper than the true slope, the estimated intercept will be below the true intercept. Similarly, if the estimated slope is much flatter than the true slope (or has a negative slope), the estimated intercept will be above the true intercept. This generates the hourglass shape.

27. It is approximately normally distributed with a mean of 2 and a standard error of 0.9. Because the histogram is centered on β_1, the estimator is an unbiased estimator.

28. Approximately 60% of the values would reject H_o because they would be outside the interval (–1.8, 1.8). Power is the probability of correctly rejecting the null hypothesis. In this case, since the null hypothesis is false, it is correct to reject it. The expected and observed results are similar because the experiment is showing that a replication experiment duplicates what regression theory tells us will happen. The only time this will not be true is if the regression assumptions are not satisfied.

29. It is approximately normally distributed with a mean of -50 and a standard error of 80. It is unbiased because the histogram of estimated intercepts is centered on β_0 (10).

30. Approximately 2% of the values would reject H_o because they would be outside the interval (–160, 160). It has very low power since the standard error is so large. Power is reduced if the standard error is large and is increased if the standard error is reduced.

31. The first histogram is positively skewed with a mean of about 0.1 and a range of 0.4. The second histogram is positively skewed with a mean of about 0.4 and range of 1. The third histogram is much less skewed with a mean of about 0.08 and a range of 0.3. Since the distribution of R^2 is not known, it can’t be used as a test statistic.

Advanced Learning Exercises

32. Results will vary. See exercise 33 for discussion.

33. The means of the estimated intercept, slope, and variance of the model do not change as the sample size increases, because these are unbiased estimators. The means of R^2 and r increase as the sample size increases, because it is easier to fit a model with more observations. The range of all the estimates is decreased as the sample size increases because estimates based on a larger sample are more stable (less variability).

34. Results will vary. See exercise 35 for discussion.

35. The means of the estimated intercept, slope, and variance of the model do not change as the range of X increases. The means of R^2 and r increase as the range of X increases (i.e., there is a better fit). In contrast, the ranges of the estimated intercept, slope, and variance of the model *decrease* as the range of the independent variable increases, while the ranges of R^2 and r *increase* as the range of X increases.

36. The means of the estimated intercept, slope, and variance of the model do not change with the range of X because the estimators are unbiased. However, the means of R^2 and r increase as the range of X increases because a better fit is achieved with the increased range. The ranges of the estimated intercept, slope, and variance of the model are reduced as the range of X increases because increasing the range of the independent variable makes the estimates more accurate. In contrast, the ranges of R^2 and r increase as the range of the independent variable increases because the larger range makes it possible for the fit to be more erratic.

37. The estimated variance of the model is distributed as a scaled chi-square. It has this distribution because the estimated variance is the sum of squared, normally distributed random variables. The histogram of the standard error is not used because its distribution is used much less frequently (scaled chi distribution).

38. The correlation coefficient is not normally distributed. However, Sir Ronald Fisher showed that a transformation of the correlation coefficient is normally distributed.