Chapter 11: Analyzing Association Between Quantitative Variables: Regression Analysis

Section 11.4: What Do We Learn from How the Data Vary Around the Regression Line?

Learning Objectives

1. Residuals and Standardized Residuals
2. Analyzing Large Standardized Residuals
3. The Residual Standard Deviation
4. Confidence Interval for μ_y
5. Prediction Interval for y
6. Prediction Interval for y vs Confidence Interval for μ_y

Learning Objective 1: Residuals and Standardized Residuals

- **A *residual* is a prediction error – the difference between an observed outcome and its predicted value**
  - **The magnitude of these residuals depends on the units of measurement for y**
- **A *standardized version of the residual* does not depend on the units**

Learning Objective 1: Standardized Residuals

- **Standardized residual:**

$$\frac{(y-\hat{y})}{se(y-\hat{y})}$$

- **The *se* formula is complex, so we rely on software to find it**
- **A standardized residual indicates how many standard errors a residual falls from 0**
- **If the relationship is truly linear and the standardized residuals have approximately a bell-shaped distribution, observations with standardized residuals larger than 3 in absolute value often represent outliers**

Learning Objective 1:
Example: Detecting an Underachieving College Student

- **Data was collected on a sample of 59 students at the University of Georgia**
- **Two of the variables were:**
  - **CGPA: College Grade Point Average**
  - **HSGPA: High School Grade Point Average**

5

Learning Objective 1:
Example: Detecting an Underachieving College Student

- **A regression equation was created from the data:**
  - **x: HSGPA**
  - **y: CGPA**
- **Equation:** $\hat{y} = 1.19 + 0.64x$

6

Learning Objective 1:
Example: Detecting an Underachieving College Student

MINITAB highlights observations that have standardized residuals with absolute value larger than 2:

TABLE 12.6: Observations with Large Standardized Residuals in Student GPA Regression Analysis, as Reported by MINITAB

Obs	HSGPA	CGPA	Fit	Residual	St Resid ← *standardized residuals*
14	3.30	2.60	3.29	−0.69	−2.26R
28	3.80	2.98	3.61	−0.63	−2.01R
59	3.60	2.50	3.48	−0.98	−3.14R

R denotes an observation with a large standardized residual.

7

Learning Objective 1:
Example: Detecting an Underachieving College Student

- **Consider the reported standardized residual of -3.14**
  - **This indicates that the residual is 3.14 standard errors below 0**
  - **This student's actual college GPA is quite far below what the regression line predicts**

8

Learning Objective 2: Analyzing Large Standardized Residuals

- **Does it fall well away from the linear trend that the other points follow?**
- **Does it have too much influence on the results?**
- **Note: Some large standardized residuals may occur just because of ordinary random variability-even if the model is perfect, we'd expect about 5% of the standardized residuals to have absolute values > 2 by chance.**

Learning Objective 2: Histogram of Residuals

- **A histogram of residuals or standardized residuals is a good way of detecting unusual observations**
- **A histogram is also a good way of checking the assumption that the conditional distribution of y at each x value is normal**
  - **Look for a bell-shaped histogram**

Learning Objective 2: Histogram of Residuals

- **Suppose the histogram is not bell-shaped:**
  - **The distribution of the residuals is not normal**

However….

  - **Two-sided inferences about the slope parameter still work quite well**
  - **The t- inferences are *robust***

Learning Objective 3: The Residual Standard Deviation

- **For statistical inference, the regression model assumes that the conditional distribution of y at a fixed value of x is normal, with the same standard deviation at each x**
- **This standard deviation, denoted by σ, refers to the variability of y values for all subjects with the same x value**

Learning Objective 3: The Residual Standard Deviation

- **The estimate of σ, obtained from the data, is:**

$$s = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

Learning Objective 3: Example: How Variable are the Athletes' Strengths?

- **From MINITAB output, we obtain s, the residual standard deviation of y:**

$$s = \sqrt{\frac{3522.8}{55}} = 8.0$$

- **For any given x value, we estimate the mean y value using the regression equation and we estimate the standard deviation using s: s = 8.0**

Learning Objective 4: Confidence Interval for μ_y

- **We estimate μ_y, the population mean of y at a given value of x by:** $\hat{y} = a + bx$
- **We can construct a 95% confidence interval for μ_y using:** $\hat{y} \pm t_{.025}(se\ of\ \hat{y})$

where the *t*-score has *df*=n-2

Learning Objective 5: Prediction Interval for y

- **The estimate $\hat{y} = a + bx$ for the *mean of y* at a fixed value of x is also a prediction for an *individual outcome y* at the fixed value of x**
- **Most regression software will form this interval within which an outcome y is likely to fall**

$$\hat{y} \pm 2s$$

where s is the residual standard deviation

Learning Objective 6:
Prediction Interval for y vs Confidence Interval for μ_y

- **The prediction interval for y is an inference about where individual observations fall**
  - **Use a *prediction interval for y* if you want to predict where a single observation on y will fall for a particular x value**

Learning Objective 6:
Prediction Interval for y vs Confidence Interval for μ_y

- **The confidence interval for μ_y is an inference about where a population mean falls**
- **Use a *confidence interval for μ_y* if you want to estimate the mean of y for all individuals having a particular x value**

$$\hat{y} \pm 2\left(s/\sqrt{n}\right)$$

where s is the residual standard deviation

Learning Objective 6:
Prediction Interval for y vs Confidence Interval for μ_y

- **Note that the prediction interval is wider than the confidence interval - you can estimate a population mean more precisely than you can predict a single observation**
- **Caution: in order for these intervals to be valid, the true relationship must be close to linear with about the same variability of y-values at each fixed x-value**

Learning Objective 6:
Example: Predicting Maximum Bench Press and Estimating its Mean

TABLE 12.7: MINITAB Output for Confidence Interval (CI) and Prediction Interval (PI) on Maximum Bench Press for Athletes Who Do 11 60-Pound Bench Presses before Fatigue

```
Predicted Values for New Observations
New Obs    Fit   SE Fit      95% CI            95% PI
   1      79.94   1.06   (77.81, 82.06)   (63.76, 96.12)
Values of Predictors for New Observations
New Obs   BP_60
   1       11.0
```