SUMMARY: Review of Residuals from Chapter 3

- Each observation has a residual. Some are positive, some are negative, some may be zero, and their average equals 0.
- In the scatterplot, a residual is the vertical distance between the data point and the regression line. The smaller the distance, the better the prediction. (See margin figure.)
- We can summarize how near the regression line the data points fall by

$$\text{sum of squared residuals} = \Sigma(\text{residual})^2 = \Sigma(y - \hat{y})^2.$$

- The regression line has a smaller sum of squared residuals than any other line. It is called the **least squares** line, because of this property.

SUMMARY: The Regression Line Connects the Estimated Means of *y* at the Various *x* Values

$\hat{y} = a + bx$ describes the relationship between x and the estimated means of y at the various values of x.

SUMMARY: Regression Model

A **regression model** describes how the population mean μ_y of each conditional distribution for the response variable depends on the value x of the explanatory variable. A **straight-line regression model** uses the line $\mu_y = \alpha + \beta x$ to connect the means. The model also has a parameter σ that describes variability of observations around the mean of y at each x value.

SUMMARY: Properties of the Correlation, *r*

- The correlation r has the same sign as the slope b. Thus, $r > 0$ when the points in the scatterplot have an upward trend and $r < 0$ when the points have a downward trend.
- The correlation r always falls between -1 and $+1$, that is, $-1 \leq r \leq +1$.
- The larger the absolute value of r, the stronger the linear association, with $r = \pm 1$ when the data points all fall exactly on the regression line.

SUMMARY: Relationship of Correlation and Slope

If the data have the same amount of variability for each variable, with $s_x = s_y$, then, $r = b$: The correlation and the slope are the same. (See margin figure.)

- The correlation r does not depend on the units of measurement.
- The correlation represents the value that the slope equals if the two variables have the same standard deviation.

SUMMARY: Properties of r^2

- Since $-1 \leq r \leq 1$, r^2 falls between 0 and 1.
- $r^2 = 1$ when $\Sigma(y - \hat{y})^2 = 0$, which happens only when all the data points fall exactly on the regression line. There is then no prediction error using x to predict y (that is, $y = \hat{y}$ for each observation). This corresponds to $r = \pm 1$. (See margin figure.)
- $r^2 = 0$ when $\Sigma(y - \hat{y})^2 = \Sigma(y - \bar{y})^2$. This happens when the slope $b = 0$, in which case each $\hat{y} = \bar{y}$. The regression line and $\bar{y}$ then give the same predictions.
- The closer r^2 is to 1, the stronger the linear association: The more effective the regression equation $\hat{y} = a + bx$ then is compared to $\bar{y}$ in predicting y.

SUMMARY: Correlation *r* and Its Square r^2

Both the correlation r and its square r^2 describe the strength of association. They have different interpretations. The correlation falls between -1 and $+1$. It represents the slope of the regression line when x and y have equal standard deviations. It governs the extent of "regression toward the mean." The r^2 measure falls between 0 and 1 (or 0% and 100% when reported by software in percentage terms). It summarizes the reduction in sum of squared errors in predicting y using the regression line instead of using the mean of y.

SUMMARY: Basic Assumption for Using Regression Line for Description

The population means of y at different values of x have a straight-line relationship with x, that is, $\mu_y = \alpha + \beta x$.

SUMMARY: Extra Assumptions for Using Regression to Make Statistical Inference

- The data were gathered using randomization, such as random sampling or a randomized experiment.
- The population values of y at each value of x follow a normal distribution, with the same standard deviation at each x value.

SUMMARY: Steps of Two-Sided Significance Test About a Population Slope β

1. **Assumptions:** (1) Population satisfies regression line $\mu_y = \alpha + \beta x$, (2) data gathered using randomization, (3) population y values at each x value have normal distribution, with same standard deviation at each x value.
2. **Hypotheses:** H_0: $\beta = 0$, H_a: $\beta \neq 0$
3. **Test statistic:** $t = (b - 0)/se$, where software supplies sample slope b and its se.
4. **P-value:** Two-tail probability of t test statistic value more extreme than observed, using t distribution with $df = n - 2$.
5. **Conclusions:** Interpret P-value in context. If a decision is needed, reject H_0 if P-value $\leq$ significance level (such as 0.05).

SUMMARY: Prediction Interval for *y* and Confidence Interval for μ_y at Fixed Value of *x*

For large samples with an x value equal to or close to the mean of x,

- The 95% **prediction interval** for y is approximately $\hat{y} \pm 2s$.
- The 95% **confidence interval** for μ_y is approximately

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

where s is the residual standard deviation. Software uses *exact* formulas. We show these *approximate* formulas here merely to give a sense of what these intervals do. *In practice, use software.*

SUMMARY: Exponential Regression Model

An **exponential regression model** has the formula

$$\mu_y = \alpha\beta^x$$

for the mean μ_y of y at a given value of x, where α and β are parameters.