Chapter 12

Regression Analysis

How *x* and *y* are related

Regression Analysis

The first step of a *regression analysis* is to identify the response and explanatory variables.

- We use *y* to denote the *response variable.*
- We use *x* to denote the *explanatory variable.*

The Scatterplot

The first step in answering the question of association is to look at the data.

A scatterplot is a graphical display of the relationship between the response variable (y) and the explanatory variable (x).

Example: The Strength Study

An experiment was designed to measure the strength of female athletes.

The goal of the experiment was to determine if there is an association between the maximum number of pounds that each individual athlete could bench press and the number of 60-pound bench presses that athlete could do.

57 high school female athletes participated in the study.

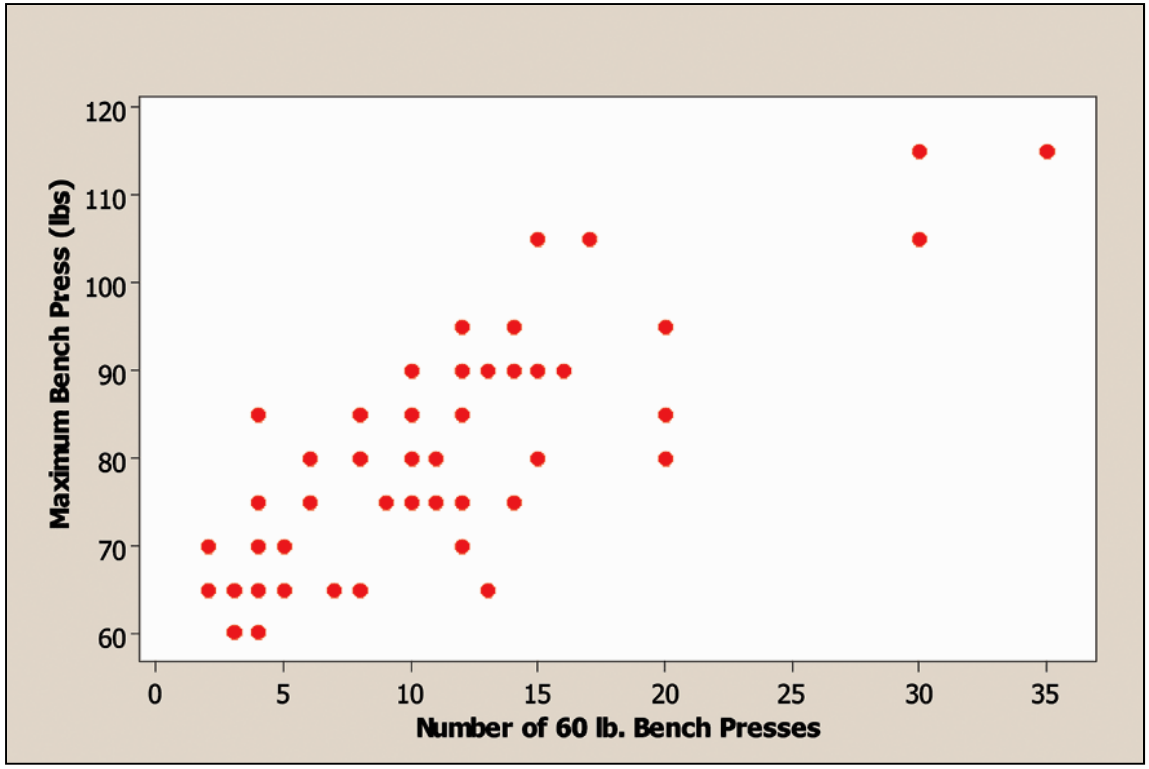
The data consisted of the following variables:

- x: the number of 60-pound bench presses an athlete could do.

  x: mean = 11.0, st. deviation = 7.1
- y: maximum bench press.

  y: mean = 79.9 lbs, st. dev. = 13.3 lbs

Example: The Strength Study



When the scatterplot shows a linear trend, a straight line can be fitted through the data points to describe that trend.

The regression line is: $\hat{y} = a + bx$

$\hat{y}$ is the predicted value of the response variable y,

a is the y-intercept and b is the slope.

Example: Regression Line Predicting Maximum Bench Press

```
The regression equation is BP = 63.5 + 1.49 BP_60

Predictor     Coef  SE Coef     T      P

Constant    63.537    1.956 32.48  0.000

BP_60       1.4911   0.1497  9.96  0.000
```

y=Maximum Bench Press (BP) and x =Number of 60-Pound Bench Presses (BP_60)

The MINITAB output shows the following regression equation:

BP = 63.5 + 1.49 (BP_60)

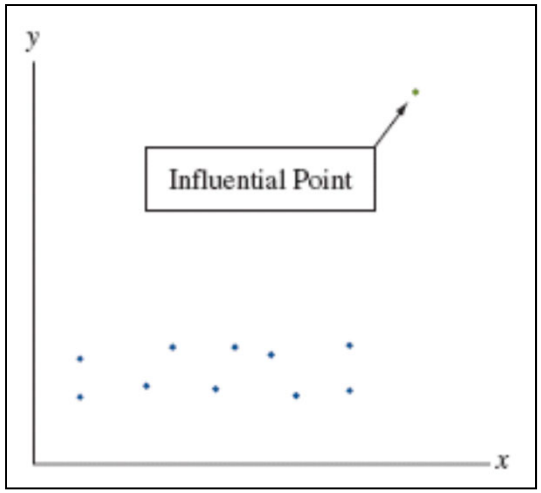
The y-intercept is 63.5 and the slope is 1.49.

The slope of 1.49 tells us that predicted maximum bench press increases by about 1.5 pounds for every additional 60-pound bench press an athlete can do.

Outliers

Check for *outliers* by plotting the data.

The regression line can be pulled toward an outlier and away from the general trend of points.



Influential Points

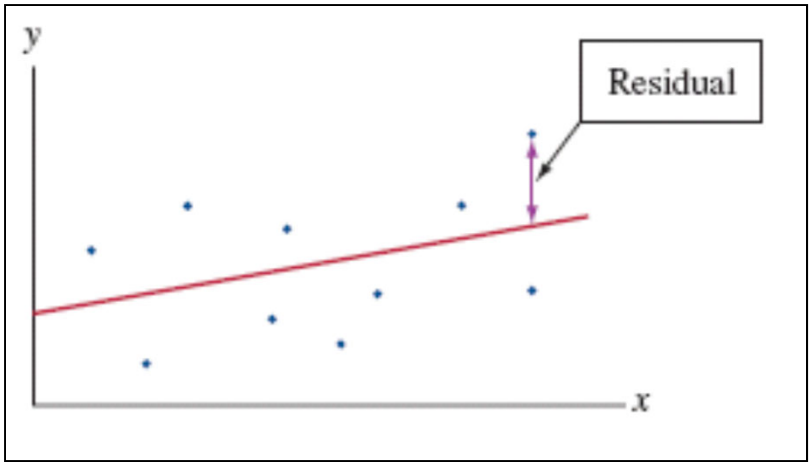
An observation can be influential in affecting the regression line when one or more of two things happen:

- Its x value is low or high compared to the rest of the data.
- It does not fall in the straight-line pattern that the rest of the data have

Residuals are Prediction Errors

The regression equation is often called a **prediction equation.**

The difference $y - \hat{y}$ between an observed outcome and its predicted value (vertical distance between the data point and the regression line) is the prediction error, called a **residual.**



Each observation has a residual.

The smaller the distance, the better the prediction.

Review of Residuals

We can summarize how near the regression line the data points fall by

$$sum\ of\ \ squared\ residuals =$$

$$\sum (residuals)^2 = \sum (y - \hat{y})^2$$

The regression line has the smallest sum of squared residuals and is called the **least squares** line.

Variability about the Line

At a given value of x, the equation: $\hat{y} = a + bx$

- Predicts a single value of the response variable.
- But… we should not expect all subjects at that value of x to have the same value of y because variability occurs in the y values.

At each fixed value of *x*, variability occurs in the *y* values around their mean, μ_y .

The probability distribution of *y* values at a fixed value of *x* is a **conditional distribution.**

At each value of *x*, there is a conditional distribution of *y* values.

An additional parameter σ describes the standard deviation of each conditional distribution.

The Population Regression Equation

The population regression equation describes the relationship in the population between x and the means of y.

The equation is denoted by: $\mu_y = \alpha + \beta x$

α is a population *y*-intercept and

β is a population slope.

These are parameters, so in practice their values are unknown.

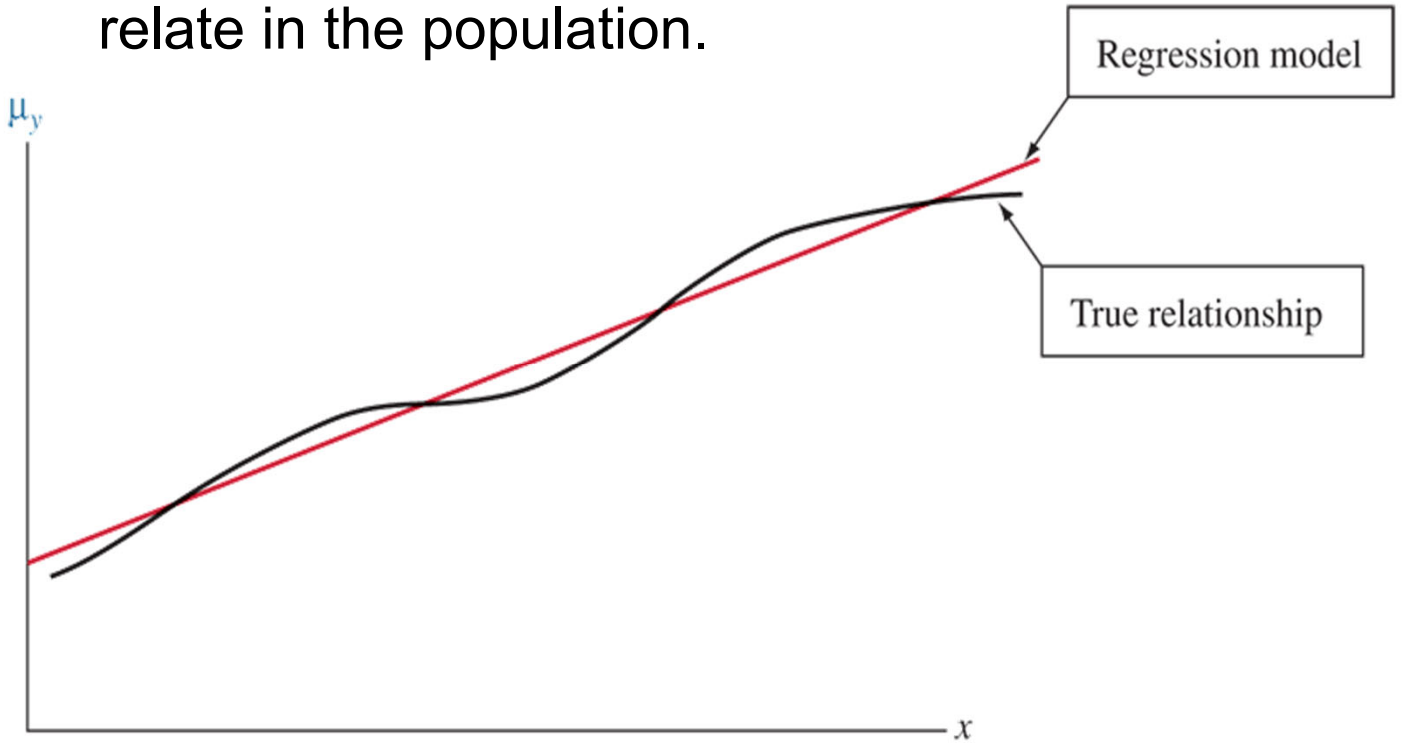
In practice we estimate the population regression equation using the prediction equation for the sample data.

The Regression Model

The **population regression equation** merely approximates the actual relationship between *x* and the population means of *y.*

It is a **model.**

- A **model** is a simple approximation for how variables relate in the population.



The Regression Model $\mu_y = \alpha + \beta x$ for the Means of y Is a Simple Approximation for the True Relationship.

A Statistical Model

A statistical model never holds exactly in practice.

It is merely an approximation for reality.

Even though it does not describe reality exactly, a model is useful if the true relationship is close to what the model predicts.

Chapter 12

Regression Analysis

(recall) Describe Strength of Association

Properties of the Correlation, *r*

The correlation, denoted by *r*, describes *linear association.*

The correlation '*r*' has the same sign as the slope '*b*'.

The correlation '*r*' always falls between -1 and +1.

The larger the absolute value of *r*, the stronger the linear association.

Correlation and Slope

We can't use the slope to describe the strength of the association between two variables because the slope's numerical value depends on the units of measurement.

The correlation does not depend on units of measurement.

The correlation and the slope are related in the following way:

$$r = b \frac{s_x}{s_y}$$

Example: Predicting Strength

For the female athlete strength study:

- x: number of 60-pound bench presses
- y: maximum bench press
- x: mean = 11.0, st.dev.=7.1
- y: mean= 79.9 lbs., st.dev. = 13.3 lbs.

Regression equation: $\hat{y} = 63.5 + 1.49x$

$$r = b\left(\frac{s_x}{s_y}\right) = 1.49\left(\frac{7.1}{13.3}\right) = 0.80$$

The variables have a **strong, positive** association.

The Squared Correlation

Another way to describe the strength of association refers to how close predictions for y tend to be to observed y values.

The variables are strongly associated if you can predict y much better by substituting x values into the prediction equation than by merely using the sample mean $\bar{y}$ and ignoring x.

The Squared Correlation

Another way to describe the strength of association refers to how close predictions for y tend to be to observed y values.

The variables are strongly associated if you can predict y much better by substituting x values into the prediction equation than by merely using the sample mean $\bar{y}$ and ignoring x.

When a **strong linear association exists**, the regression equation predictions tend to be much better than the predictions using $\bar{y}$.

We measure the proportional reduction in error and call it, r^2.

Chapter 12

Regression Analysis

Make Inferences About the Association

Descriptive and Inferential Parts of Regression

The sample regression equation, r, and r^2 are descriptive parts of a regression analysis.

The inferential parts of regression use the tools of confidence intervals and significance tests to provide inference about the regression equation, the correlation and r-squared in the population of interest.

Assumptions for Regression Analysis

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:

$$\mu_y = \alpha + \beta x$$

- This can be verified with a scatterplot.

Extra assumptions for using regression to make statistical inference:

- The data were gathered using randomization.
- The population values of y at each value of x follow a normal distribution, with the same standard deviation at each x value.

Assumptions for Regression Analysis

Models, such as the regression model, merely approximate the true relationship between the variables.

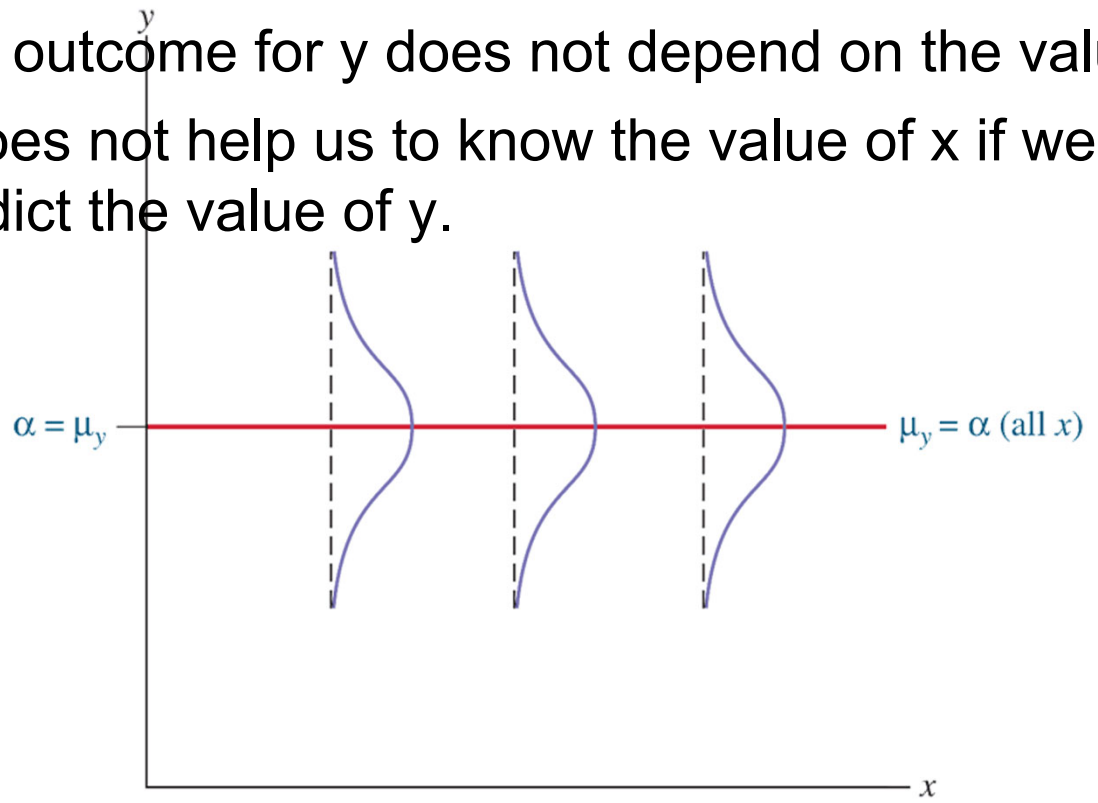
A relationship will not be exactly linear, with exactly normal distributions for y at each x and with exactly the same standard deviation of y values at each x value.

Testing Independence between Quantitative Variables

Suppose that the slope β of the regression line equals 0

Then…

- The mean of y is identical at each x value.
- The two variables, x and y, are statistically independent.
- The outcome for y does not depend on the value of x.
- It does not help us to know the value of x if we want to predict the value of y.



Testing Independence between Quantitative Variables

1. **Hypotheses:** $H_0 : \beta = 0, H_a : \beta \neq 0$
2. **Assumptions:**
   - The population satisfies regression line:
   $$\mu_y = \alpha + \beta x$$
   - Data obtained using randomization
   - The population values of y at each value of x follow a normal distribution, with the same standard deviation at each x value.

3. **Test statistic:** $t = \dfrac{b - 0}{se}$
   - Software supplies sample slope b and its se

Testing Independence between Quantitative Variables

4. *P*-value: Two-tail probability of *t* test statistic value more extreme than observed:

Use *t* distribution with $df = n - 2$

5. Conclusions: Interpret *P*-value in context. If decision needed, reject H_0 if *P*-value $\leq$ significance level.

Example: 60-Pound Strength and Bench Presses

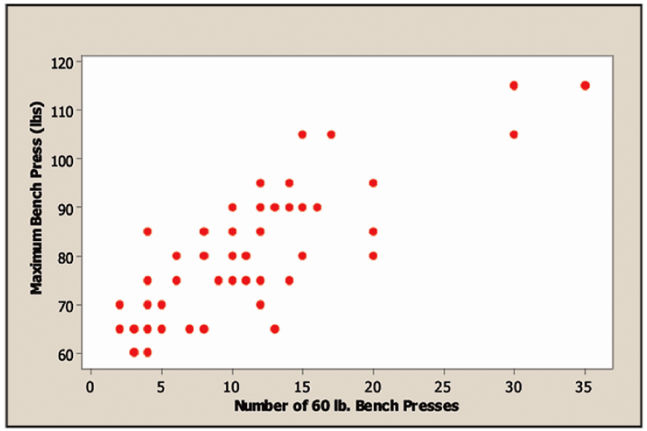
The regression equation is BP = 63.5 + 1.49 BP_60

Predictor	Coef	SE Coef	T	P
Constant	63.537	1.956	32.48	0.000
BP_60	1.4911	0.150	9.96	0.000

R-Sq = 64.3%

Example: 60-Pound Strength and Bench Presses

1. **Hypotheses:** $H_0 : \beta = 0, H_a : \beta \neq 0$
2. **Assumptions:**
   - A scatterplot: linear trend.
   - The scatter of points have a similar spread at different x values.
   - The sample was a convenience sample, not a random sample, so this is a concern.



3. Test statistic: $t = \frac{b - 0}{se} = \frac{(1.49 - 0)}{0.150} = 9.96$

4. P-value: 0.000

5. Conclusion: An association exists between the number of 60-pound bench presses and maximum bench press.

A Confidence Interval for β

A small *P*-value in the significance test of $H_0 : \beta = 0$ suggests that the population regression line has a nonzero slope.

To learn how far the slope β falls from 0, we construct a confidence interval:

$$b \pm t_{.025}(se) \text{ with } df = n - 2$$

Example: Estimating the Slope for Predicting Maximum Bench Press

A small *P*-value in the significance test of $H_0 : \beta = 0$ suggests that the population regression line has a nonzero slope.

To learn how far the slope β falls from 0, we construct a confidence interval: $b \pm t_{.025}(se)$ with $df = n - 2$

Construct a 95% confidence interval for β.

$$1.49 \pm 2.00(0.150) \; which \; is:$$

$$1.49 \pm 0.30 \; or \; (1.2, 1.8)$$

Based on a 95% CI, we can conclude, on average, the maximum bench press increases by between 1.2 and 1.8 pounds for each additional 60-pound bench press that an athlete can do.

Example: Estimating the Slope for Predicting Maximum Bench Press

Let's estimate the effect of a 10-unit increase in x:

- Since the 95% CI for β is (1.2, 1.8), the 95% CI for 10β is (12, 18).

- On the average, we infer that the maximum bench press increases by at least 12 pounds and at most 18 pounds, for an increase of 10 in the number of 60-pound bench presses.

Chapter 12

Regression Analysis

How the Data Vary Around the Regression Line

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.

Standardized Residuals

Standardized residual = $\dfrac{(y - \hat{y})}{se(y - \hat{y})}$

The *se* formula is complex (use 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.

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

Example: Detecting an Underachieving College Student

A regression equation was created from the data:

- x: HSGPA
- y: CGPA

Equation: $\hat{y} = 1.19 + 0.64\,x$

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.

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.

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.

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.

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.

The estimate of σ, obtained from the data, is called the **residual standard deviation**:

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

Example: Variability of 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* = 8.0.

Confidence Interval for μ_y

We can 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 $df = n - 2$

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

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.

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

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.

Maximum Bench Press and Estimating its Mean

```
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
```

For all female high school athletes who can do 11 sixty-pound bench presses, we estimate the **mean** of their maximum bench press values falls between 78 and 82 pounds.

For all female high school athletes who can do 11 sixty-pound bench presses, we predict that 95% of **them** have maximum bench press values between 64 and 96 pounds.