# Estimation and Prediction with the Regression Model

**Major goals in using the regression model:**
**(1) Determining the linear relationship between *Y* and *X* (accomplished through inferences about $\beta_1$)**

**(2) Estimating the mean value of *Y*, denoted E(*Y*), for a particular value of *X*.**
**Example: Among all people with drug amount 3.5 ~~mg~~ percent, what is the estimated mean reaction time?**

**(3) Predicting the value of *Y* for a particular value of *X*.**
**Example: For a "new" individual having drug amount 3.5 ~~mg~~ %, what is the predicted reaction time?**

- **The point estimate for these last two quantities is the same; it is:** the $\hat{Y}$ corresponding to that particular X value.

**Example:** $\hat{Y} = -0.1 + 0.7X$

For $X = 3.5$, $\hat{Y} = -0.1 + 0.7(3.5) = 2.35$ secs

So for $X = 3.5$ percent, the estimated $E(Y)$ is 2.35 and the predicted Y value is 2.35.

- **However, the variability associated with these point estimates is very different.**

- **Which quantity has more variability, a single Y-value or the mean of many Y-values?**

A single Y-value will have more variability.

**This is seen in the following formulas:**

**100(1 – α)% <u>Confidence Interval</u> for the population mean value of $Y$ at $X = x_p$:**

$$\hat{y} \pm t_{\alpha/2} (s) \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

**where $t_{\alpha/2}$ based on $n - 2$ d.f.**

**100(1 – α)% <u>Prediction Interval</u> for ~~the~~ an individual new value of $Y$ at $X = x_p$:**

$$\hat{y} \pm t_{\alpha/2} (s) \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

**where $t_{\alpha/2}$ based on $n - 2$ d.f.**

**The extra "1" inside the square root shows the prediction interval is wider than the CI, although they have the same center.**

**Note: A "Prediction Interval" attempts to contain a random quantity, while a confidence interval attempts to contain a (fixed) parameter value.**

$$E(Y) = \beta_0 + \beta_1 X \qquad \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

**The variability in our estimate of E(*Y*) reflects the fact that we are merely estimating the unknown $\beta_0$ and $\beta_1$.**

**The variability in our prediction of the new *Y* includes that variability, <u>plus</u> the natural variation in the Y-values.**

**Example (drug/reaction time data):**
**95% CI for E(*Y*) with *X* = 3.5:**

Recall: $SS_{xx} = 10$, $s = \sqrt{MSE} = .606$.

$$\hat{Y} = -0.1 + 0.7(3.5) = 2.35$$

$$n = 5, \quad \bar{X} = 3$$

$$\text{t-table: } t_{\alpha/2} = t_{.025}\ (\overset{n-2}{3}\ \text{d.f.}) = 3.182$$

$$2.35 \pm 3.182\,(.606)\sqrt{\frac{1}{5} + \frac{(3.5-3)^2}{10}}$$

$$2.35 \pm 3.182(.606)(.47434)$$

$$\Rightarrow 2.35 \pm .91467 \Rightarrow \boxed{(1.44,\ 3.26)}$$

→ We are 95% confident that the mean reaction time for people with drug amount 3.5 percent is between 1.44 and 3.26 seconds.

**95% PI for a new *Y* having *X* = 3.5:**

$$2.35 \pm (3.182)(.606)\sqrt{1 + \frac{1}{5} + \frac{(3.5-3)^2}{10}}$$

$$2.35 \pm (3.182)(.606)(1.1068)$$

$$2.35 \pm 2.134 \Rightarrow \boxed{(0.216,\ 4.484)}$$

With 0.95 probability, we predict that the reaction time for a new person with drug amount 3.5 percent will be between 0.216 and 4.484 seconds.

# Complete Regression Example

$Y$ = damage (in thousands of dollars)
$X$ = distance from station (in miles)

- Scatter plot shows a linear relationship between damage and distance.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\hat{y} = 10.278 + 4.919 X$$

Interpreting estimated slope: We estimate that for each one-mile increase in distance from station, the predicted damage increases by 4.919 thousand dollars.

- Residual plots show model assumptions are reasonable.

Test: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$

test statistic $t = 12.53$, P-value near 0.

$\Rightarrow$ Reject $H_0$. Conclude distance is linearly related to damage (distance is a useful predictor of damage).

95% CI for $\beta_1$: $(4.07, 5.77)$

Correlation: $r = .961 \Rightarrow$ Strong, positive linear relationship between damage and distance.

Interpreting $r^2 = .9235$: 92.35% of the sample variation in damage can be explained by its linear relationship with distance from station.

- Consider a hypothetical fire 3.5 miles from fire station:
- For a distance of $X = 3.5$ miles from station: predicted damage is $\hat{Y} = 27.496$ thousand dollars.

95% PI for new $Y$ with $X = 3.5$

$\Rightarrow (22.32, 32.67)$

- With probability 0.95, the fire damage for a house 3.5 miles from station will be between 22.32 and 32.67 thousand dollars.

95% CI for $E(Y)$ when $X = 3.5$

$\Rightarrow (26.19, 28.80)$

With 95% confidence, the mean damage for houses 3.5 miles from station is between 26.19 and 28.80 thousand dollars.