Name: ____

1. The Leaning Tower of Pisa is an architectural wonder. Engineers concerned about the towerâĂŹs stability have done extensive studies of its increasing tilt. Measurements of the lean of the tower over time provide much useful information. Data was collected over a period of 13 consecutive years (x), from 1975–1987. The engineers measured the amount of lean (y) in meters as the distance in meters between where a point on the tower would be if the tower were straight and where it actually is.

   Here are the summary statistics: $n = 13$, $\bar{x} = 1981$, $s_x = 3.894$, $\bar{y} = 2.96937$, $s_y = 0.00365$, $r = 0.994$, $\hat{\sigma} = 0.00042$.

   (a) (6 points) Estimate the parameters of the simple linear regression model. Provide units for the estimated parameters.

   Solution: The slope is $\hat{\beta}_2 = rs_y/s_x = (0.994)(0.00365)/(3.894) \doteq 0.000932$ meters per year and the intercept is $\bar{y} - \hat{\beta}_2\bar{x} = 2.96937 - (0.0009317)(1981) \doteq 1.12364$ meters. Units for the slope are units of y per units of x. Units for the intercept are units of y.

   (b) (3 points) Use the regression line to predict the amount of lean in 2009.

   Solution: Plugging into the equation yields

   $$\hat{y} = 1.12364 + (0.0009317)(2009) \doteq 2.99546 \; .$$

   (c) (3 points) The formula for the standard error of a prediction for $\hat{y}$ at x is

   $$\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

   and the 0.975 quantile from an appropriate t distribution is 2.201. Use this formula to find a 95% prediction interval for the amount of lean in 2009.

   Solution: Note that

   $$\sum_{i=1}^n (x_i - \bar{x})^2 = (n-1)s_x^2 \doteq 182$$

   so a numerical estimate of the standard error is

   $$0.00042\sqrt{\frac{1}{13} + \frac{(2009-1981)^2}{182}} \doteq 0.00088$$

   and a 95% prediction interval is then

   $$2.99546 \pm (2.201)(0.00088), \quad 2.99546 \pm (0.00194), \quad \text{or} \quad (2.99352, 2.99739)$$

   (d) (2 points) How many degrees of freedom does the t quantile from the previous problem have?

   Solution: $n - 2 = 11$.

   (e) (3 points) Does the prediction interval from part (c) properly account for all uncertainty in the prediction? Briefly explain.

   Solution: The prediction interval accounts for uncertainty in the estimated regression line and uncertainty associated with individual variation, *but it does not account for extrapolation*. Namely, even if a linear model is appropriate between time and the amount of leaning from 1975–1987, there is uncertainty in the assumption that this linear trend continues all the way to 2009. This model uncertainty is **not** accounted for in the prediction interval. Recall the example from lecture where estimates of my son's height at age 0 or much older than when the data was collected were poor because the linear relationship did not extend outside the range of the collected data.

(f) (3 points) Briefly respond to the validity of the following statement.

> The observed correlation coefficient $r = 0.994$ is very close to one which provides very strong evidence that a linear model is adequate and that a more general model where $\mathsf{E}(Y \mid X)$ was a curve would not significantly improve the fit to the data.

Solution: The statement is not valid. The correlation coefficient indicates that a line with a positive slope fits the data much better than a horizontal line, but r says nothing about the relative appropriateness of a nonlinear model to a linear model. Recall the identical point I made in lecture when discussing the model for my son's growth over time, both when we first encountered regression and when we reviewed with the practice exam.