AP Statistics Name: _____
Chapter 12 – Inference for Regression Date: _____ Period: ___

Topic: Inference for Regression

We have looked at inference for means and proportions, for one and two samples. We have looked at inference for categorical data. Once of those tests was to see if there was an association between two categorical variables. In this unit, we look at inference procedures about the relationship between two *quantitative* variables measured on the same individuals.

Suppose we wanted to answer questions like:

"Is there really a relationship between arm-span and height in adults, or could the pattern we see in the scatterplot plausibly happen by chance?"

"In the population of adults, how much will the predicted height change for each increase of 1 inch in arm span? What is the margin of error for this estimate?"

Questions like these involve **regression**.
We will revisit the Normal and *t* distributions for our test statistic.

Procedures for inference still follow the State, Plan, Do, Conclude method. Some of the details are different, particularly the conditions for inference. We also learn some techniques to consider if some of the conditions are not met.

If you do not remember what we did with regression, then it would be a good idea to re-read Chapter 3.

Vocabulary you should recall from Chapter 3:

explanatory variable	Variable that may help explain or predict changes in a response variable. (Graphed on the horizontal axis.)
response variable	Variable that measures an outcome of a study. (Graphed on the vertical axis.)
scatterplot	Plot that shows the relationship between two quantitative variables measured on the same individuals.
correlation *r*	Measures the direction and strength of the linear relationship between two quantitative variables. (On formula sheet.)
direction	the direction of the association of two quantitative variables: positive, negative or something else (describe)
form	the form of the association of two quantitative variables: linear, curved, or something else (describe)
strength	the strength of the association of two quantitative variables: described such as: weak, moderate, strong or maybe none
outlier	Observation that lies outside the overall pattern of the other observations.
positive association	When above-average values of one variable to accompany above-average values of the other and also of below-average values to occur together.
negative association	When above-average values of one variable tend to accompany below-average values of the other.
regression line	Line that describes how a response variable y changes as an explanatory variable x changes.
predicted value	$\hat{y}$ (read "y hat") is the predicted value of the response variable y for a given value of the explanatory variable x.
slope	In a regression line, the amount by which y is predicted to change when x increases by one unit.
y-intercept	In a regression line, the predicted value of y when x = 0.
extrapolation	Use of a regression line for prediction far outside the interval of values of the explanatory variable x used to obtain the line. Such predictions are often not accurate.
residual	Difference between an observed value of the response variable and the value predicted by the regression line: $y - \hat{y}$
least-squares regression line (LSRL)	The line that makes the sum of the squared vertical distances of the data points from the line as small as possible.
residual plot	Scatterplot of the residuals against the explanatory variable. Residual plots help us assess whether a linear model is appropriate.
standard deviation of the residuals (s)	This value gives the approximate size of a "typical" prediction error (residual).
coefficient of determination: r^2	Fraction of the variation in the values of y that is accounted for by the least-squares regression line of y on x.
influential observations	An observation is influential for a statistical calculation if removing it would markedly change the result of the calculation.

New Vocabulary introduced in Chapter 12:

$\mu_y = \alpha + \beta x$	True regression line based on the entire population of data.
$\sigma_b = \frac{\sigma}{\sigma_x\sqrt{n}}$	The standard deviation of the sampling distribution of b.
$SE_b = \frac{s}{s_x\sqrt{n-1}}$	The standard error of the sampling distribution of b. This approximates σ_b using sample data.
L-I-N-E-R	• Linear – the linear model $\mu_y = \alpha + \beta x$ exists • Independent – the observations are independent • Normal – all the y distributions are Normal for each x value • Equal SD – all the y distributions have the same σ. • Random – the data come from a random sample or a randomized experiment
t interval for slope	Confidence interval for β, the true slope of the regression line
t test for slope	Significance test for β, the true slope of the regression line.
ρ (rho)	The true correlation r measuring the direction and strength of the linear relationship between the two quantitative variables.
transforming data	Transforming data can sometimes change curved relationships between two quantitative variables into linear relationships.
power model	A curved relation of the form $y = ax^P$. Can be transformed to achieve linearity by any of: (x^P, y) , $\left(x, \sqrt[P]{y}\right)$, (lnx, lny).
logarithm	In stat, we generally use natural logarithms: $e^A = B \iff lnB = A$
exponential model	A curved relation of the form $y = ab^x$. Can be transformed to achieve linearity by (x, lny).

Summary of Inference for β, True Slope of the Regression Line

Test of Significance	
State	$H_0: \beta = \beta_0$ $H_a: \beta < \beta_0$ or $H_a: \beta \neq \beta_0$ or $H_a: \beta > \beta_0$ Where β is the true slope of the regression line of <response variable> on <explanatory variable>. $\alpha = \cdots$ (select a significance level)
Plan	If conditions for inference are met, conduct a *t* test for linear regression. Linear: Check to see that a linear model fits the data; linear is a good fit if the scatterplot of the residuals show a patternless scatter. Independent: Check that the observations are independent. When sampling without replacement, check the 10% condition. Normal: Plot the residuals to make sure there is no strong skew or outliers; a probability plot should be approximately linear. Equal SD: Scatterplot of the residuals vs. the explanatory variable should show about the same (vertical) spread for every level of *x*. Random: Check that the data come from a random sample or a randomized experiment.
Do	the test statistic is *t* with $n-2$ degrees of freedom: $t = \frac{b - \beta_0}{SE_b} \sim t(n-2)$ The P-value is $P(T < t)$ or $2P(T > \lvert t \rvert)$ or $P(T > t)$ depending on the direction of your alternative hypothesis
Conclude	If P-Value is less than alpha (significance level), then we have evidence in favor of the alternative hypothesis. Otherwise, we do NOT have evidence in favor of the alternative hypothesis.
Confidence Interval	If conditions for inference are met, a CI% *t* interval for β, the true slope of the regression line of <response variable> on <explanatory variable> is given by: $b \pm t^* SE_b$ where the critical value t^* comes from a *t* distribution with $n-2$ degrees of freedom.