CHAPTER 11

Study Questions

1. Why do we call the technique used to find the regression line the method of least squares?

2. What is an interpretation for the slope of the regression line?

3. What implications are there if the slope is equal to zero?

4. What are some applications for linear regression?

5. What does the correlation coefficient measure?

6. What does it mean when we say two variables are positively correlated?

7. Does a high correlation imply causality?

8. What are two hypotheses testing methods that may be used to test if an independent variable is useful in determining the value of a dependent variable?

9. What is the difference between a confidence interval and a prediction interval?

10. What assumption about the random error is used to conduct hypothesis test?

10. What is a residual?

11. How are residual plots used in linear regression?

Linear Regression

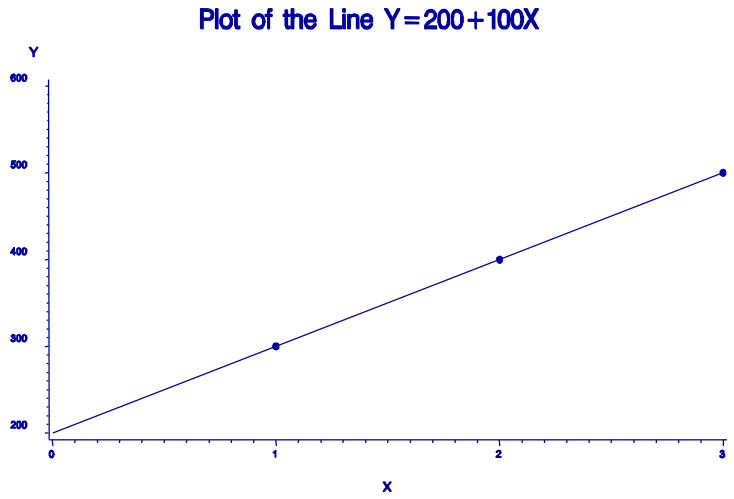
Linear regression: continuous Y and continuous X.

Deterministic Line
The equation for a line is given by:

$$Y = \beta_0 + \beta_1 X$$

where : Y = the dependent variable (variable being predicted)
X = the independent variable (variable used for the prediction)
β_0 = the y-intercept
β_1 = the slope of the line

Below is data and the line for Y= 200+100X. Notice that the data points and the line match exactly.



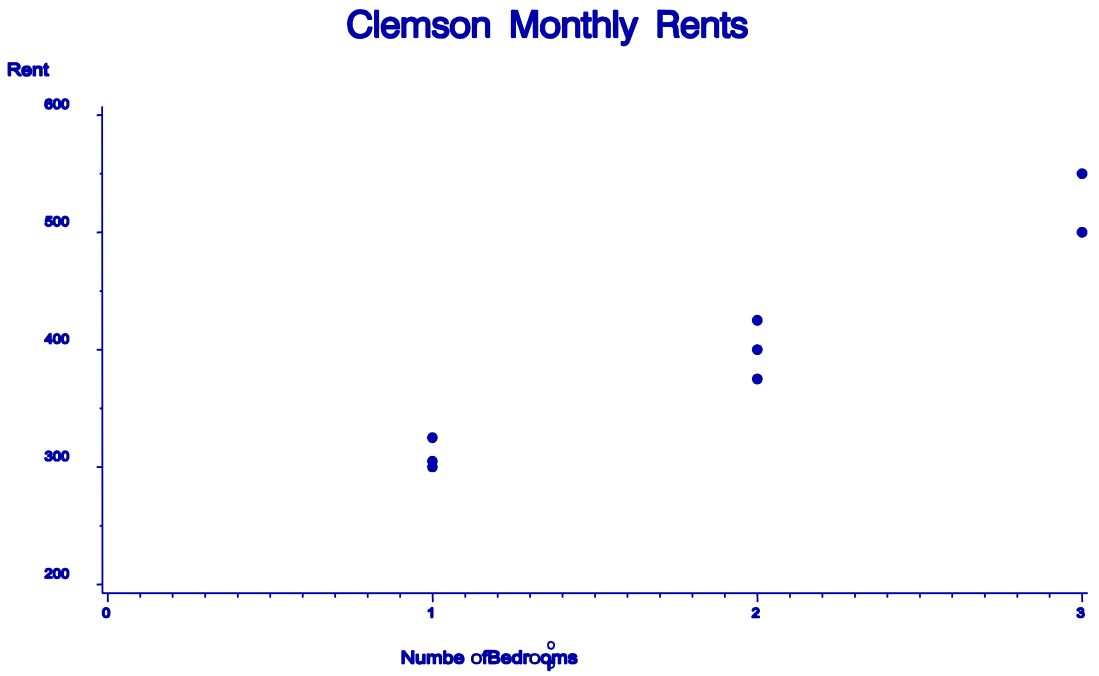
What is the value of the y-intercept?

What is the value for the slope of the line?

Example: In the table below is a random sample of rents taken from a rental pricing survey. All of the apartments have air conditioning, are unfurnished and have no dishwasher.

Apartment	Number of Bedrooms (X)	Monthly Rent (Y)
Charleston Ave.	1	300
Clarendon Drive	1	305
College Street Apartments	1	325
Brookdale Apartments	2	375
Porter House Apartments	2	400
Robin Hill Apartments	2	425
Robin Hill Apartments	3	500
Berkley Drive	3	550

Below is a graph of the data for the Clemson rent example. There appears to be a linear relationship between the number of bedrooms and the monthly rent. We would like to fit a line through the data so that it can be used to predict the expected rent for different size apartments. The question is, what is the "best" line to fit to this data?



Probabilistic Model

The probabilistic linear model is $Y = \beta_0 + \beta_1 X + \varepsilon$

where
- Y = the dependent variable
- X = the independent variable
- β_0 = the y-intercept
- β_1 = the slope of the line
- ε = the random error term.

The random error term takes into account all unknown factors that are not included in the model.

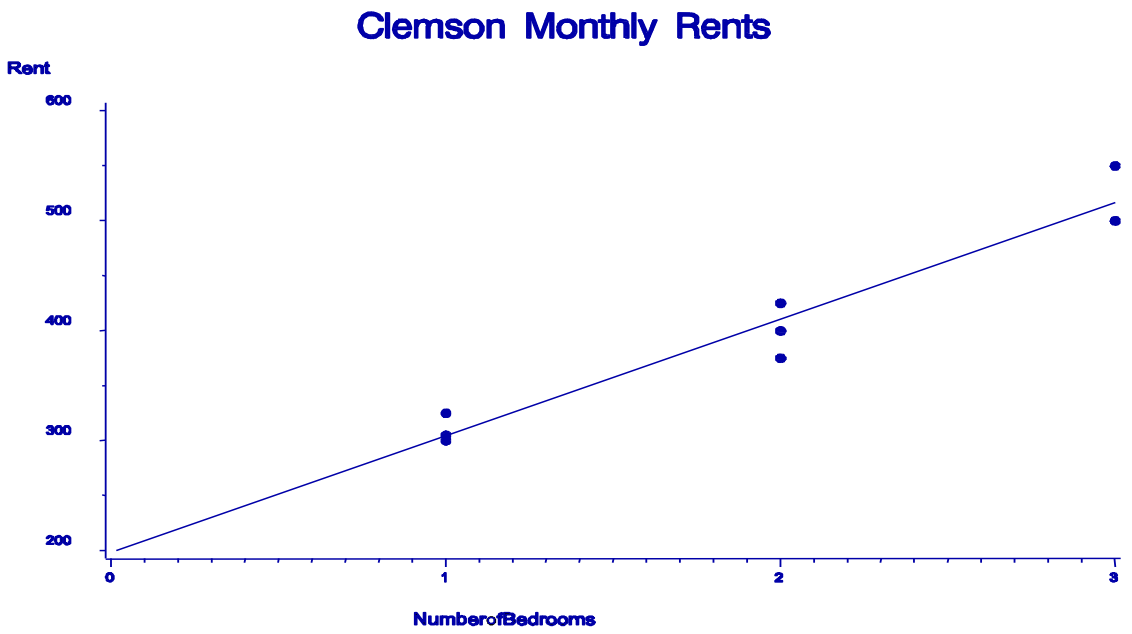
Assumptions about ε

1. The ε's are independent.
2. The random error ε for each level of X has a mean of 0 and a variance of σ^2.
3. The variance σ^2 is the same for each level of X.
4. The random error has a normal distribution for each level of X.

If assumption 1 through 3 are satisfied, then the expected value of Y for a given value of X is unbiased.
If assumption 4 is satisfied, then confidence intervals and hypothesis testing may be done using the t-distribution.

The expected value of Y for a given value of x is denoted by

$$E(Y) = \beta_0 + \beta_1 x$$



Method of Least Squares

When we draw a line through the data, we call it the **prediction line** and write it as

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

The difference between the actual y-value and the predicted y-value (y hat) is called the **observed error**. So the observed error when X is equal to x_i is denoted by

$$e_i = y_i - \hat{y}_i$$

The observed error is also known as the deviation. One way to find the prediction line would be to minimize the sum of the squared deviations. The name for this method is called the **Method of Least Squares**. Below are the **least squares estimators** for the fitted line.

Estimators for the Slope (β_1) and Y-intercept (β_0):

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Estimator for σ_ε^2

$$s_e^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{n - 2}$$

Apartment	Number of Bedrooms (X)	Monthly Rent (Y)	X^2	XY
Charleston Ave.	1	300	1	300
Clarendon Drive	1	305	1	305
College Street Apartments	1	325	1	325
Brookdale Apartments	2	375	4	750
Porter House Apartments	2	400	4	800
Robin Hill Apartments	2	425	4	850
Robin Hill Apartments	3	500	9	1500
Berkley Drive	3	550	9	1650
Sum	15	3180	33	6480

Find the estimates of β_1 and β_0, then write the prediction equation.

Predict the monthly rent for a four-bedroom apartment.

Predict the monthly rent for an efficiency apartment (no bedrooms).

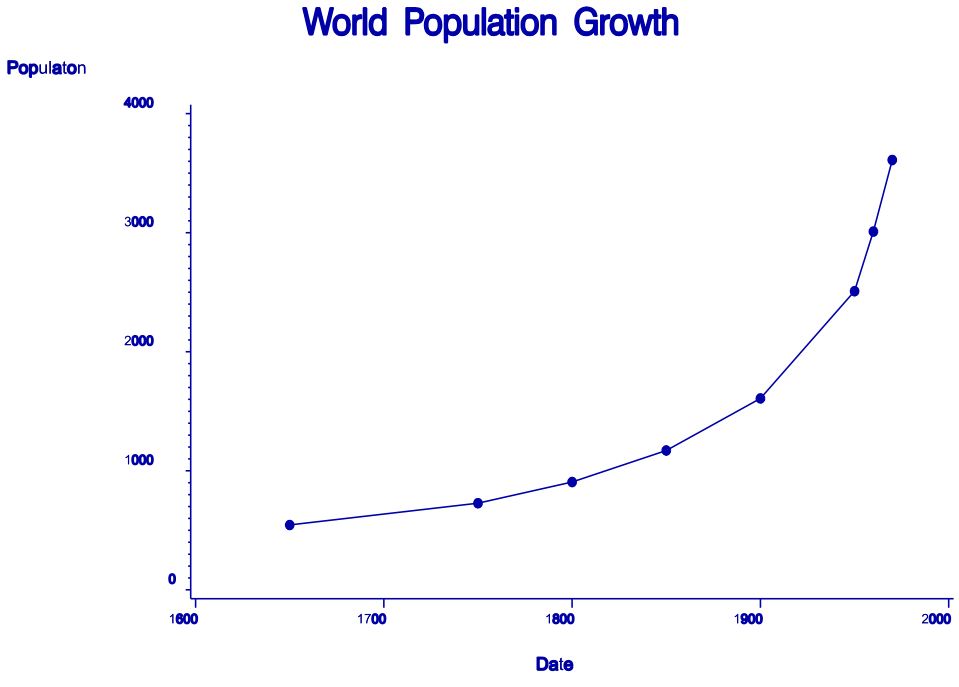
Transformations to Linearize Data

There may be situations where the relationship between the Y and X variables is not linear. In some cases, it may still be possible to establish a linear relationship if we transform the Y and/or the X variable.

Example:

Below is a table of world population growth. When we plot the population growth (Y) vs. the date (X), we see that there is not a linear relationship between the two variables. Is there some transformation that will "linearize" the data?

Date	World Population (in millions)
1650	545
1750	728
1800	906
1850	1171
1900	1608
1950	2509
1960	3010
1970	3611



What transformation or transformations would be good candidates to linearize the relationship between date and population?

We can transform date to “date ^2”.

Least Squares Line

Prediction of the size of the population for 1980.

Prediction of the size of the population for 1990.

Prediction of the size of the population for 2000.

Decomposing the Sum of Squares about the Mean

$$\underset{i=1}{\overset{n}{\sum}}\left(y_i - \bar{y}\right)^2 = \underset{i=1}{\overset{n}{\sum}}\left(\hat{y}_i - \bar{y}\right)^2 + \underset{i=1}{\overset{n}{\sum}}\left(y_i - \hat{y}_i\right)^2$$

Sum of Squares about the Mean | *Sum of Squares Due to regression* | *Sum of Squares for Error*

Decomposition of the Sum of Squares about the Mean
for the Clemson Rent Example

Obs.	y_i	$\left(y_i - \bar{y}\right)^2$	$\left(\hat{y}_i - \bar{y}\right)^2$	$\left(y_i - \hat{y}_i\right)^2$
1	300	9506.25	8628.55	21.25
2	305	8556.25	8628.55	0.15
3	325	5256.25	8628.55	415.75
4	375	506.25	175.83	1278.78
5	400	6.25	175.83	115.78
6	425	756.25	175.83	202.78
7	500	10506.25	14261.14	286.29
8	550	23256.25	14261.14	1094.29
SUM		58350.00	54935.42	3415.07

$$\text{SST} = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = S_{yy} =$$

$$\text{SSR} = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 = \frac{S_{xy}^2}{S_{xx}} =$$

$$\text{SSE} = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right) =$$

The proportion (percent) of variation explained by the regression is

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

which is called the coefficient of determination. (Note: for simple regression $R^2 = r^2$).

Calculations for the Clemson Rent Problem

Apartment	Number of Bedrooms (X)	Monthly Rent (Y)	X^2	Y^2	XY
Charleston Ave.	1	300	1	90000	300
Clarendon Drive	1	305	1	93025	305
College Street Apartments	1	325	1	105625	325
Brookdale Apartments	2	375	4	140625	750
Porter House Apartments	2	400	4	160000	800
Robin Hill Apartments	2	425	4	180625	850
Robin Hill Apartments	3	500	9	250000	1500
Berkley Drive	3	550	9	302500	1650
Sum	15	3180	33	1322400	6480

Calculate:

$$s_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 4.875$$

$$s_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 58350$$

$$s_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = 517.5$$

Sum of Squares about the Mean $\quad SSTotal = $ $s_{yy} = 58350$

Sum of Squares Due to Regression $\quad SSREG = \frac{s_{xy}^2}{s_{xx}} = \frac{(517.5)^2}{4.875} = 54935$

Sum of Squares for Error $\quad SSE = s_{yy} - \frac{s_{xy}^2}{s_{xx}} = 58350 - 54935 = 3415$