Chapter 1 – Linear Regression with One Predictor Variable

Regression Analysis

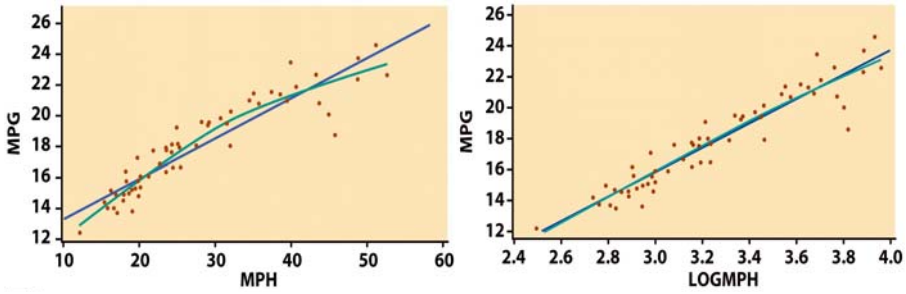
- A "tool" used to serve three purposes (1) Description; (2) Control; (3) Prediction
- Should **not** be used to imply causality
- **Always** need to consider scope of the model

Objectives

- Statistical model
- Estimating the regression parameters
- Properties of Least Squares Estimates
- Estimation of the error variance

Example: Relationship between speed and fuel efficiency Computers in some vehicles calculate various quantities related to the vehicle's performance. One of these is the fuel efficiency, or gas mileage, expressed as miles per gallon (mgp). Another is the average speed in miles per hour (mph). For one vehicle equipped in this way, mpg and mph were recorded each time the gas tank was filled, and the computer was then reset. How does the speed at which the vehicle is driven affect the fuel efficiency?

- Goals:
  - To characterize fuel efficiency relationship
  - To predict fuel efficiency for a speed of 30 mph
- Response/Dependent variable: mpg (Y)
- Explanatory/Independent variable: mph (X)
- Is there a relationship between Y and X?
- Is there a linear relationship between Y and X?
- Is there a linear relationship between Y and log(X) where log(X)=LOGPMH?



1.3 Statistical Model

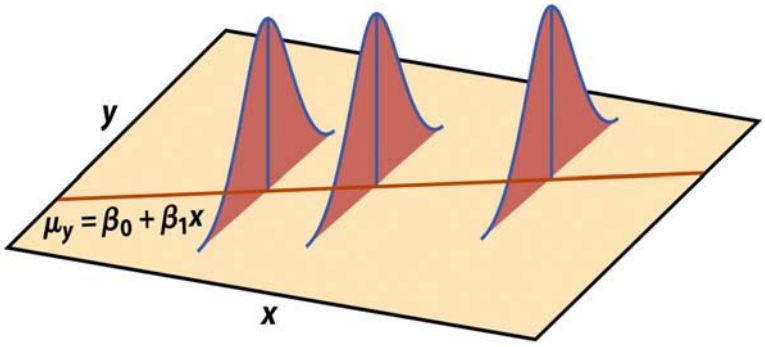
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1,\ldots,n$$

where:

- Y_i is the value of the response for the i^{th} trial
- β_0, β_1 are parameters
- X_i is a known constant, the value of the predictor variable for the i^{th} trial
- $\beta_0 + \beta_1 X_i$ is the mean response when X= X_i
- ε_i is a random error term, such that:

$$E\{\varepsilon_i\} = 0 \quad \sigma^2\{\varepsilon_i\} = \sigma^2 \quad \sigma\{\varepsilon_i, \varepsilon_j\} = 0 \quad \forall\, i, j \ni i \neq j$$

The last point states that the random errors are independent (uncorrelated), with mean 0, and variance σ^2.
This regression model is: (1) simple; (2) linear in the parameters; (3) linear in the predictor variables.



Important Features of Model:

The response Y_i is the sum of two components: (1) the constant term $\beta_0 + \beta_1 X_i$; and (2) the random term ε_i. Therefore Y_i is a random variable with

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \qquad \sigma^2\{Y_i\} = \sigma^2 \quad \sigma\{Y_i, Y_j\} = 0$$

Meaning of Regression Parameters

The parameters β_0 and β_1 are called **regression coefficients**. Thus, β_0 represents the mean response when $X = 0$ (assuming that is reasonable level of X), and is referred to as the ***Y*-intercept**. When the scope of the model does not cover X=0, β_0 does not have any particular meaning. Also, β_1 represent the change in the mean response as X increases by 1 unit, and is called the **slope**.

1.6 Least Squares Estimation of Model Parameters

In practice, the parameters β_0 and β_1 are unknown and must be estimated. One widely used criterion is to minimize the sum of squared deviations (or errors):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad \Rightarrow \quad \varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))^2$$

This is done by calculus, by taking the partial derivatives of Q with respect to β_0 and β_1 and setting each equation to 0. The values of β_0 and β_1 that set these equations to 0 are the **least squares estimates** and are labeled b_0 and b_1.

First, take the partial derivates of Q with respect to β_0 and β_1:

$$\frac{\partial Q}{\partial \beta_0} = 2\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))(-1) \qquad (1)$$

$$\frac{\partial Q}{\partial \beta_0} = 2\sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_i))(-X_i) \qquad (2)$$

Next, set these 2 equations to 0, replacing β_0 and β_1 with b_0 and b_1 since these are the values that minimize the error sum of squares:

$$-2\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i \qquad (1a)$$

$$-2\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) X_i = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \qquad (2a)$$

These two equations are referred to as the **normal equations** (although, note that we have said nothing YET, about normally distributed data).

Solving these two equations yields:

$$b_1 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} Y_i = \sum_{i=1}^{n} k_i Y_i$$

$$b_0 = \overline{Y} - b_1 \overline{X} = \sum_{i=1}^{n} \left[\frac{1}{n} - \overline{X} k_i \right] Y_i = \sum_{i=1}^{n} l_i Y_i$$

where k_i and l_i are constants, and Y_i is a random variable with mean and variance given above:

$$k_i = \frac{X_i - \overline{X}}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$l_i = \frac{1}{n} - \overline{X}k_i = \frac{1}{n} - \frac{\overline{X}(X_i - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Properties of Least Squares Estimates

Gauss-Markov Theorem: These least squares estimates

(1) Are unbiased;
(2) Have minimum variance among all unbiased linear estimators.

The **fitted regression line**, also known as the **prediction equation** is:

$$\hat{Y} = b_0 + b_1 X$$

The **fitted values** for the individual observations aye obtained by plugging in the corresponding level of the predictor variable (X_i) into the fitted equation.

Extensions of Markov-Gauss Theorem:

- $E(\hat{Y}_i) = E(Y_i)$
- $\hat{Y}_i$ has the minimum variance among all linear estimators

The **residuals** are the vertical distances between the **observed values** (Y_i) and their **fitted values** ($\hat{Y}_i$), and are denoted as e_i .

$$\hat{Y}_i = b_0 + b_1 X_i \qquad e_i = Y_i - \hat{Y}_i$$

Properties of Fitted Regression Line:

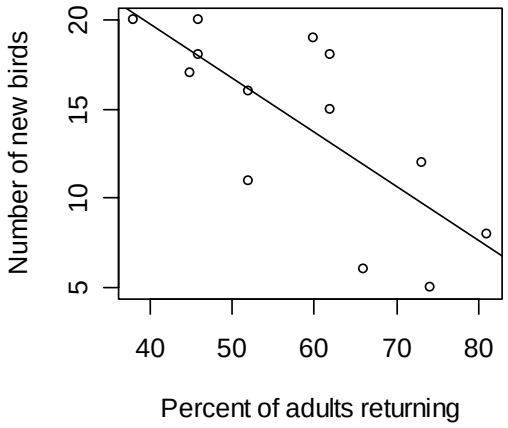
- $\sum_{i=1}^{n} e_i = 0$ The residuals sum to 0
- $\sum_{i=1}^{n} e_i^2$ is a minimum
- $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$ The sum of the observed values equals the sum of the fitted values
- $\sum_{i=1}^{n} X_i e_i = 0$ The sum of the weighted (by X) residuals is 0

- $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$ The sum of the weighted (by $\hat{Y}$) residuals is 0
- The regression line goes through the point $(\overline{X}, \overline{Y})$

These can be derived via their definitions and the normal equations.

Example: Bird colonies. One of nature's patterns connects the percent of adult birds in a colony that return from the previous year and the number of new adults that join the colony. Here are data for 13 colonies of sparrowhawks:

	X_i	Y_i	$X_i - \overline{X}$	$Y_i - \overline{Y}$	$(X_i - \overline{X})(Y_i - \overline{Y})$	$(X_i - \overline{X})^2$	$(Y_i - \overline{Y})^2$
	74	5	15.77	-9.23	-145.56	248.67	85.21
	66	6	7.77	-8.23	-63.95	60.36	67.75
	81	8	22.77	-6.23	-141.87	518.44	38.82
	52	11	-6.23	-3.23	20.13	38.82	10.44
	73	12	14.77	-2.23	-32.95	218.13	4.98
	62	15	3.77	0.77	2.90	14.21	0.59
	52	16	-6.23	1.77	-11.02	38.82	3.13
	45	17	-13.23	2.77	-36.64	175.05	7.67
	62	18	3.77	3.77	14.21	14.21	14.21
	46	18	-12.23	3.77	-46.10	149.59	14.21
	60	19	1.77	4.77	8.44	3.13	22.75
	46	20	-12.23	5.77	-70.56	149.59	33.28
	38	20	-20.23	5.77	-116.72	409.28	33.28
Total	757	185	0.00	0.00	-619.69	2038.31	336.31
Mean	58.23	14.23					



1) Obtain the least squares estimates of β_0 and β_1.
2) State the regression function.
3) Predict how many new adult birds will join another colony, to which 60% of the adults from previous year return.

1.7 **Estimation of the Error Variance**

Note that for a random variable, its variance is the expected value of the squared deviation from the mean. That is, for a random variable W, with mean μ_W its variance is:

$$\sigma^2\{W\} = E\{(W - \mu_W)^2\}$$

For the simple linear regression model, the errors have mean 0, and variance σ^2. This means that for the actual observed values Y_i, their mean and variance are as follows:

$$E\{Y_i\} = \beta_0 + \beta_1 X_i \qquad \sigma^2\{Y_i\} = E\{(Y_i - (\beta_0 + \beta_1 X_i))^2\} = \sigma^2$$

First, we replace the unknown mean $\beta_0 + \beta_1 X_i$ with its fitted value $\hat{Y}_i = b_0 + b_1 X_i$, then we take the "average" squared distance from the observed values to their fitted values. We divide the sum of squared errors by *n*-2 (2 degrees of freedom are lost due to the parameter estimates) to obtain an unbiased estimate of σ^2 (recall how you computed a sample variance when sampling from a single population).

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

Common notation is to label the numerator as the **error sum of squares (SSE)** or **residual sum of squares**.

$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Also, the estimated variance is referred to as the **error (or residual) mean square (MSE)**.

$$MSE = s^2 = \frac{SSE}{n-2}$$

To obtain an estimate of the standard deviation (which is in the units of the data), we take the square root of the error mean square. $s = \sqrt{MSE}$.

1.8 Normal Error Regression Model

If we add further that the random errors follow a normal distribution, then the response variable also has a normal distribution, with mean and variance given above. The notation, we will use for the errors, and the data is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, \ldots, n$$

- Y_i is the value of the response for the i^{th} trial
- β_0, β_1 are parameters
- X_i is a known constant, the value of the predictor variable for the i^{th} trial
- ε_i are independent $N(0, \sigma^2)$ (**NEW: Normally distributed)**

$E\{\varepsilon_i\} = 0 \quad \sigma^2\{\varepsilon_i\} = \sigma^2 \quad \sigma\{\varepsilon_i, \varepsilon_j\} = 0 \quad \forall\, i, j \ni i \neq j$

The density function for the i^{th} observation is:

$$f_i = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma} \right)^2 \right]$$

The likelihood function is the product of the individual density functions (due to the independence assumption on the random errors).

$$L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2 \right]$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \right]$$

The values of $\beta_0, \beta_1, \sigma^2$ that maximize the likelihood function are referred to as **maximum likelihood estimators**. The MLE's are denoted as: $\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_2$. Note that the natural logarithm of the likelihood is maximized by the same values of $\beta_0, \beta_1, \sigma^2$ that maximize the likelihood function, and it's easier to work with the log likelihood function.

$$\log_e L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Taking partial derivatives with respect to $\beta_0, \beta_1, \sigma^2$ yields:

$$\frac{\partial \log L}{\partial \beta_0} = -2\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)(-1) \qquad (4)$$

$$\frac{\partial \log L}{\partial \beta_1} = -2\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)(-X_i) \qquad (5)$$

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \qquad (6)$$

Setting these three equations to 0, and placing "hats" on parameters denoting the maximum likelihood estimators, we get the following three equations:

$$\sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i \qquad (4a)$$

$$\sum_{i=1}^{n} X_i Y_i = \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \qquad (5a)$$

$$\frac{1}{\hat{\sigma}^4} \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = \frac{n}{\hat{\sigma}^2} \qquad (6a)$$

From equations 4*a* and 5*a*, we see that the maximum likelihood estimators are the same as the least squares estimators (these are the normal equations). However, from equation 6*a*, we obtain the maximum likelihood estimator for the error variance as:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2}{n} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n}$$

This estimator is biased downward. We will use the unbiased estimator $s^2 = MSE$ throughout this course to estimate the error variance.

About Normal Error Model

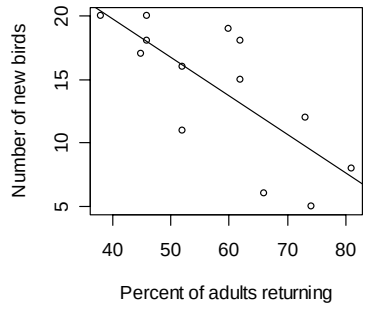
- Normal error assumption greatly simplifies the theory of analysis
- Sampling distributions used to construct confidence intervals / perform hypothesis tests follow known distributions (e.g., *t*, *F*)
- While not always true in practice, most inference only sensitive to large departures from normality

Bird colonies example using R:

```
Return <- c(74,66,81,52,73,62,52,45,62,46,60,46,38)
New <-c(5,6,8,11,12,15,16,17,18,18,19,20,20)

# plot the scatter plot
plot(Return, New, xlab="Percent of adults returning", ylab="Number of
new birds")
fit <- lm(New ~ Return)       # fit the least squars regression line
abline(fit)                   # add the regression line to the scatter
plot
summary(fit)

fit$residuals                 # residuals
fit$fitted.values             # fitted values
?lm                           # help page for lm
help(lm)                      # help page for lm
```



Program Output:

```
Call:
lm(formula = New ~ Return)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8687 -1.2532  0.0508  2.0508  5.3071

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.93426    4.83762   6.601 3.86e-05 ***
Return      -0.30402    0.08122  -3.743  0.00325 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.667 on 11 degrees of freedom
Multiple R-Squared: 0.5602,     Adjusted R-squared: 0.5202
F-statistic: 14.01 on 1 and 11 DF,  p-value: 0.003248
```