Chapter 9 - Simple linear regression models

Introduction

1. Linear regression analysis is the core of statistical modeling.
2. Applications of regression are numerous and occur in almost every field. Its uses include:

   Data description and summarization

   Parameter estimation: Ohm's law: $V = RI$

   Prediction and estimation

   Control
3. There are many types of linear models:

   Simple linear regression model: $y = \beta_0 + \beta_1 x + \varepsilon$

   y – response (dependent) variable

   x – predictor (regressor)

   (β_0, β_1) – regression coef.

   ε – error term

   Multiple linear regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... \beta_k x_k + \varepsilon$

   Polynomial regression model: $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

   The above models all require ε iid normal distribution $N(0, \sigma^2)$

   Generalized linear models (GLMs): … $y = g(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k) + \varepsilon$

   For the GLMs, ε does not usually follow a normal distribution.

4. "*linear*" means that the model is linear in parameters β_0, β_1, …, not because y is a linear function of the x's.

Estimation

1.1. Model: $y = \beta_0 + \beta_1 x + \varepsilon$

β_0 is intercept, β_1 is slope

1.2. *n* observed data pairs:

i	***y***	***x***
1	y_1	x_1
2	y_2	x_2
…	…	…
n	y_n	x_n

1.3. The most important objective of regression analysis is to estimate the unknown parameters (β_0, β_1). The process is called fitting the model to the data.

1.4. *x* is considered to be measured without error, while *y* is a random variable.

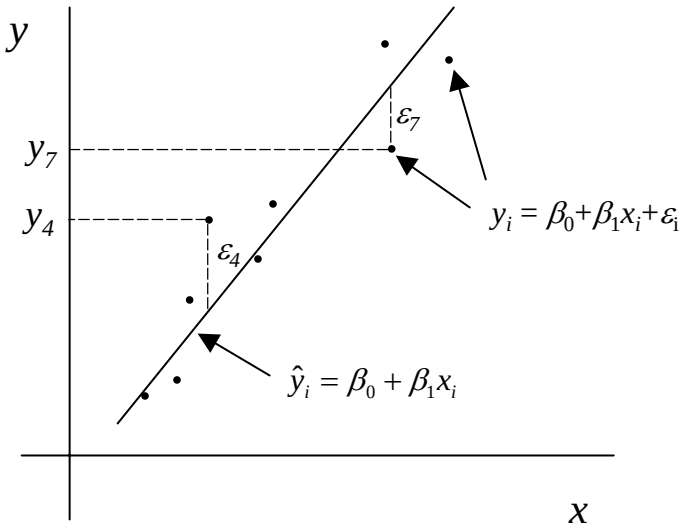
1.5. The expectation of *y* is: $E(y \mid x) = \beta_0 + \beta_1 x$

The variance of *y* is: $V(y \mid x) = V(\beta_0 + \beta_1 x + \varepsilon) = V(\varepsilon) = \sigma^2$

(Assumptions: $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2$)

1.6. Least squares estimation of parameters β_0 and β_1:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2 \rightarrow \min$$



1.7. Minimization through derivatives:

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0$$

1.8. Solve for β_0 and β_1 from (1.7):

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum y_i (x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

where $\bar{y} = \frac{1}{n}\sum y_i$, $\bar{x} = \frac{1}{n}\sum x_i$

1.9. The fitted model:

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

1.10. Residual = observed data – corresponding fitted value

$$\varepsilon_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

1.11. Example: height (y in m) versus diameter (x in cm) for 420 western hemlock trees from BC. The data object is called: **hl.dat**. y is named as **htt**, x is named as **dbh**.

R-code 1:

Step 1. Attach the data:	**attach(hl.dat)**
Step 2: Plot the data:	**plot(dbh, htt)**
Step 3. Calculate means for x and y:	**xbar=mean(dbh)**
	ybar=mean(htt)
Step 4. Calculate S_{xy} and S_{xx}:	**sxy=sum(y*(dbh-xbar))**
	sxx=sum((dbh-xbar)^2)
Step 5. Calculate β_0 and β_1:	**beta1=sxy/sxx**
	beta0=ybar-beta1*xbar
Step 6. Plot the fitted line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$:	**xx=sort(dbh)**
	yhat=beta0+beta1*xx
	lines(xx,yhat, col="red")

***R-code 2* (*use* lm)**

Step 1. One step fitting:	**hl.lm=lm(htt~dbh)**
Step 2. Plot the fitted line:	**abline(lm(hl.lm))**
Step 3. To view the regression outputs:	**summary(hl.lm)**

2. Hypothesis testing

2.1. Model assumptions:

Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, 2, \ldots, n$

The procedures developed in this section require the error term ε_i identically and independently follows the normal distribution:

$\varepsilon_i \sim N(0, \sigma^2)$,

or

$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

2.2. Properties of the LSE $\hat{\beta}_1$: BLUE (Best Linear Unbiased Estimators, "*best*" means minimum variance)

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum \frac{(x_i - \bar{x})}{S_{xx}} y_i = \sum c_i y_i$$

Unbiased: $E(\hat{\beta}_1) = \beta_1$

Variance: $V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$

Standard error of $\hat{\beta}_1$: $\frac{\sigma}{\sqrt{S_{xx}}}$

2.3. Properties of the LSE $\hat{\beta}_0$:

Unbiased: $E(\hat{\beta}_0) = \beta_0$

Variance: $V(\hat{\beta}_0) = V(\bar{y} - \hat{\beta}_1\bar{x}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)$

Standard error of $\hat{\beta}_0$: $\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$

2.4. Estimation of σ^2:

An estimate of σ^2 is required to test hypothesis and construct CI pertinent to the regression model.

Residual (error) sum of squares:

$SS_E = \sum \varepsilon_i^2 = \sum (y_i - \hat{y}_i)^2 = S_{yy} - \hat{\beta}_1 S_{xy}$

SS_E has *n*-2 df because two the 2 dfs are associated with the estimates of the two parameters $(\hat{\beta}_0, \hat{\beta}_1)$ (i.e., 2 dfs are lost due to the estimates of the two parameters).

To view residuals in R, type: **hl.lm$resid**

The unbiased estimator of σ^2 is:

$\hat{\sigma}^2 = \frac{SS_E}{n-2} = MS_E$ (*error mean square*)

$\hat{\sigma}$ is called the *standard error of regression* or *residual standard error*

2.5. Testing: H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$.

$t_0 = \frac{\hat{\beta}_1 - 0}{\sqrt{MS_E / S_{xx}}} \sim t_{n-2}$ (If H_0 is true)

Rejecting H_0 if

$|t_0| > t_{\alpha/2, n-2}$ (the upper $\alpha/2$ percentage point of the t distribution)

(Significant at α level, * for $\alpha = 0.05$; ** for $\alpha = 0.01$)

More accurate and popular expression is:

p-value = $P(t_{n-2} > |t_0|)$ {In R code, p-value = **1-pt(t_0, n-2)** }

=> Large $|t_0|$ value has small p-value, evidence against H_0.

The testing can be viewed by typing **summary(hl.lm)** in R:

> summary(hl.lm)

```
Call:
lm(formula = htt ~ dbh, data = hl.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-11.5576  -0.6854  -0.2518   0.5340  13.7582

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.26208    0.12427   10.16   <2e-16 ***
dbh          0.71762    0.01081   66.40   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 1.817 on 418 degrees of freedom
Multiple R-Squared: 0.9134,     Adjusted R-squared: 0.9132
F-statistic:  4409 on 1 and 418 degrees of freedom,     p-value:     0
```

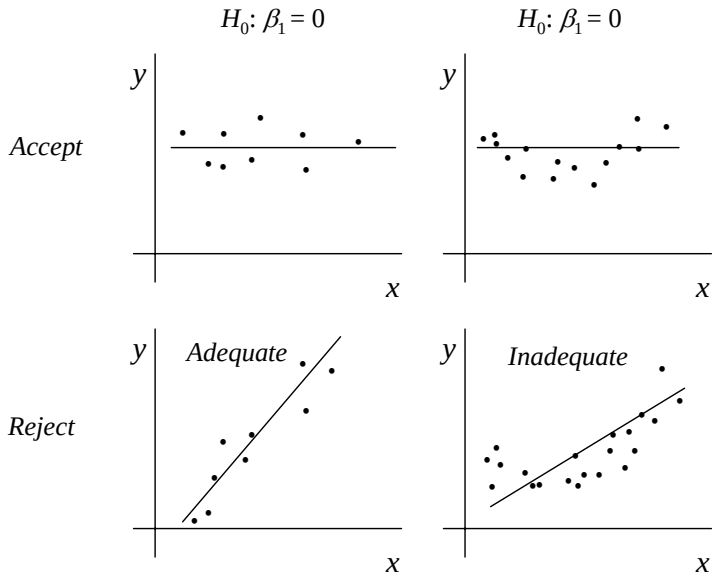
2.6. Adjusted coefficient of determination R_a^2:

$$R_a^2 = 1 - (1 - R^2)\frac{total\ df}{residual\ df} = 1 - (1 - R^2)\frac{n-1}{n-k}$$

where k is the number of parameters ($k = 2$ here).

2.7. Interpretation of the hypothesis H_0: $\beta_1 = 0$:

This is actually a test on the *significance of regression*. Failing to reject H_0 has two implications: (1) x is of little value in explaining the variation in y (x is independent of y), (2) The true relationship between x and y is not linear. In contrast, rejecting H_0 implies that (1) the linear model is adequate in explaining the variability in y, (2) x is of value in explaining y but the linear model may still not be an adequate model.



2.8. Testing: H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$. (Exercise: To formulate the test.)

$$t_0 = \frac{\hat{\beta}_0 - 0}{\sqrt{MS_E\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2} \text{ (If } H_0 \text{ is true)}$$

Rejecting H_0 if

$|t_0| > t_{\alpha/2, n-2}$ (the upper $\alpha/2$ percentage point of the t distribution)

2.9. ANOVA for testing significance of regression, equivalent to testing H_0: $\beta_1 = 0$

Identity:

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

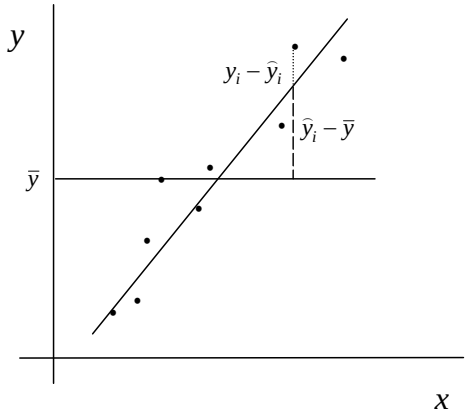
Squaring and summing both sides over all n observations:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 + 2\sum (\hat{y}_i - \bar{y})(y_i - \hat{y}_i)$$

The interaction term is zero, resulting in

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

	SS_T	=	SS_R	+	SS_E
df:	n-1		1		n-2



ANOVA Table:

Source of variation	Sum of squares	d.f.	Mean square	F_0
Regression	SS_R	1	$MS_R = SS_R/1$	MS_R/MS_E
Residual	SS_E	$n - 2$	$MS_E = SS_E/(n\text{-}2)$	
Total	SS_T	$n - 1$		

If H_0: $\beta_1 = 0$ is true, the test statistic F_0 follows the $F_{1, n-2}$ distribution. Therefore, to test the hypothesis H_0: $\beta_1 = 0$, we compare F_0 with $F_{1, n-2}$ and reject H_0 if

$F_0 > F_{a, 1, n-2}$ (Significant at α level, * for $\alpha = 0.05$; ** for for $\alpha = 0.01$)

Or to report the *p*-value:

p-value = $P(F_{1, n-2} > F_0)$ {In R-code, it is **1-pf(F_0, 1, n-2)**}

=> Large F_0 leads to small *p*-value, evidence against H_0.

ANOVA table can be produced as:

```
> anova(hl.lm)
Analysis of Variance Table

Response: htt
          Df  Sum Sq Mean Sq F value    Pr(>F)
dbh        1 14564.3 14564.3  4409.2 < 2.2e-16 ***
Residuals 418  1380.7     3.3
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1
```

2.10. Exercise: To familiarize yourself with the various statistics introduced above, suggest computing them manually as follows:

Residual sum of squares (2.4):	**sse=sum((htt - 1.26208 - 0.71762*dbh)^2)**
Residual std error (2.4):	**mse=sqrt(sse/418)**
Std error for $\hat{\beta}_1$ (2.2):	**sqrt(mse/sxx)**
Std error for $\hat{\beta}_0$ (2.3):	**sqrt(mse*(1/420+(xbar^2)/sxx))**
SS_R (2.8):	**msr=sum(1.26208 + 0.71762*dbh-ybar)^2)**
F-statistic:	**msr/mse**

How to test hypothesis H_0: $\beta_1 = 0$, based on this F-value? (Find out on the top of the previous page.)

2.11. Confidence interval for $\hat{y}$:

An important use of a regression model is to estimate the mean response $E(y|x_0)$ for a particular value x_0 of the regressor variable x. For example, we wish to estimate mean tree height at dbh of 10 cm. This height estimate is not a fixed value but a random variable whose 100(1-α)% CI is:

$$\left\{ \hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_E\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}, \quad \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_E\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \right\}$$

Exercise: Write an R code to plot the interval.

R-code:

yhat=predict(hl.lm, hl.dat, interval="confidence",level=0.95)

To add the 95% CI on y-hat:

id=order(hl.dat$dbh)
lines(hl.dat$dbh[id],yhat.lw.up[,2][id],col=2)
lines(hl.dat$dbh[id],yhat.lw.up[,3][id],col=2)

2.12. Prediction of new y:

Prediction of new observations is another important use of the regression model. Given a value of x_0, the point estimate of a new observation is simply:

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 .$$

However, the CI for the new observation is wider than the one given in 2.11 which is the CI on the mean of *y*, not a probability statement about future observations from that distribution. The CI for the new observation is:

$$\left\{ \hat{y}_0 - t_{\alpha/2,n-2}\sqrt{MS_E\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}, \quad \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{MS_E\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \right\}$$

R-code:

ynew.lw.up=predict.lm(lm.out, hl.dat, interval="prediction",level=0.95)

To add the 95% CI for the observations:

id=order(hl.dat$dbh)
lines(hl.dat$dbh[id],ynew.lw.up[,2][id],col=2)
lines(hl.dat$dbh[id],ynew.lw.up[,3][id],col=2)

How to compute the 99% CI for a new *htt* observation given *dbh* = 10 cm?

2.13. Coefficient of determination

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

SS_T is a measure of the variability in *y* without considering the effect of the regressor variable *x*, SS_E is a measure of the variability in *y* remaining after *x* has been considered. Therefore, R^2 is really the proportion of variation explained by the regressor x.

The expectation of R^2 is approximately:

$$E(R^2) \approx \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\beta}_1^2 S_{xx} + \sigma^2}$$

Clearly, R^2 depends on the range of x. R^2 will increase as the spread of the x's increases and decrease as the spread of the x's decreases. Thus, a large value of R^2 may result simply because x has been varied over an unrealistically large range. On the other hand, R^2 may be small because the range of x was too small to allow its relationship with y to be detected.

2.14. Further misconceptions about R^2

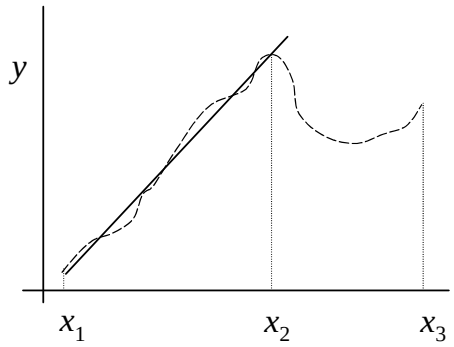
(1). In general, R^2 does not measure the magnitude of the slope of the regression line. A large value of R^2 does not imply a steep slope.

(2). R^2 does not measure the appropriateness of the linear model, for R^2 can be large even though y and x are nonlinearly related.

Take home messages

3.1. Regression models do not describe causal relationship. To establish causality, the relationship between the regressors and the response must have a basis outside the sample data (e.g., suggested by theoretical considerations). Regression analysis can aid in confirming a cause-effect relationship, but it cannot be the sole basis of such a claim.

3.2. Regression model in general is valid only over the region of the regressor variables contained in the observed data, i.e., in (x_1, x_2). Extrapolation is dangerous as illustrated below.



3.3. It is important to remember that regression analysis is part of a broader data-analysis approach to problem solving, i.e., the regression model itself may not be the primary objective of the study. We use the model to help gain insight and understanding on the system generating the data.

3.4. Data quality. Data collection is an essential aspect of regression analysis. Without representative data the regression model and conclusions drawn from it are likely to be in error.