Simple Linear Regression

1. Model and Parameter Estimation

   (a) Suppose our data consist of a collection of n pairs (x_i, y_i), where x_i is an observed value of variable X and y_i is the corresponding observation of random variable Y. The simple linear regression model

   $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

   expresses the relationship between variables X and Y. Here β_0 denotes the intercept and β_1 the slope of the regression line.

   (b) Values for β_0 and β_1 are estimated from the data by the method of least squares.

   (c) From the many straight lines that could be drawn through our data, we find the line that minimizes the sum of squared residuals, where a residual is the vertical distance between a point (x_i, y_i) and the regression line.

   (d) Values $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the estimates for β_0 and β_1 that minimize the sum of squared residuals, or error sum of squares(SSE). The estimates are called least squares estimates.

   $$SSE = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

   (e) SSE is minimized when the partial derivatives of the SSE with respect to the unknowns (β_0 and β_1) are set to zero: $\frac{\partial SSE}{\partial \beta_0} = 0$ and $\frac{\partial SSE}{\partial \beta_1} = 0$. (You need multivariable calculus [eg Math 2001] to understand the theoretical details, so we will just take this as a given.) These two conditions result in the two so-called "normal equations".

   $$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

   $$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

   (f) The two normal equations are solved simultaneously to obtain estimates of β_0 and β_1. These estimates are:

   $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

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

   Looking at the formula for $\hat{\beta}_1$, and recalling the formula for the correlation coefficient r, it is easy to see that $\hat{\beta}_1 = r s_y / s_x$.

   (g) The error variance, σ^2, is estimated as

   $$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

The following example shows the calculations as they would be carried out by hand, in gruesome detail.

eg: To study the effect of ozone pollution on soybean yield, data were collected at four ozone dose levels and the resulting soybean seed yield monitored. Ozone dose levels (in ppm)were reported as the average ozone concentration during the growing season. Soybean yield was reported in grams per plant.

X Ozone(ppm)	Y Yield (gm/plant)
.02	242
.07	237
.11	231
.15	201

- Estimated values for β_0 and β_1 are now computed from the data

X	Y	X^2	Y^2	XY
.02	242	.0004	58564	4.84
.07	237	.0049	56169	16.59
.11	231	.0121	53361	25.41
.15	201	.0225	40401	30.15

- Column sums: $\sum x_i = .35$, $\sum y_i = 911$, $\sum x_i^2 = .0399$, $\sum y_i^2 = 208,495$, and $\sum x_i y_i = 76.99$
- Means: $\bar{x} = .0875$ and $\bar{y} = 227.95$
- Intermediate terms:

$$SS_{xx} = \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - \frac{(\sum x_i)^2}{n} = .0399 - \frac{(.35)^2}{4} = .009275$$

$$SS_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 76.99 - \frac{.35(911)}{4} = -2.7225$$

- $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = -293.531$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 227.95 - (-293.531)(.0875) = 253.434$

(h) the least squares regression equation which characterizes the linear relationship between soybean yield and ozone dose is

$$\hat{y}_i = 253.434 - 293.531 x_i$$

(i) The error variance, σ^2, is estimated as MSE.

(j) Residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 * x_i)$

x_i	y_i	$\hat{y}_i$	$\hat{\epsilon}_i = y_i - \hat{y}_i$
.02	242	247.563	-5.563
.07	237	232.887	4.113
.11	231	221.146	9.854
.15	201	209.404	-8.404

(k) **Residual Sum of Squares** (In regression problems, the error sum of squares is also known as the residual sum of squares).

$$SSE = \sum \hat{\epsilon}_i^2 = (-5.563)^2 + (4.113)^2 + (9.854)^2 + (-8.404)^2 = 215.59$$

(l) Mean Squared Error: $MSE = \frac{SSE}{(n-2)} = 107.80$

```
x=c(.02,.07,.11,.15)
y=c(242,237,231,201)
SXX=sum((x-mean(x))^2)
SXY=sum((x-mean(x))*(y-mean(y)))
SYY=sum((y-mean(y))^2)
b1=SXY/SXX
b0=mean(y)-b1*mean(x)
yp=b0+b1*x
resids=y-yp
SSE=sum(resids^2)
SST=SYY
SSR=SST-SSE
SS=c(SSR,SSE,SST)
n=length(y)
df=c(1,n-2,n-1)
MS=SS/df
cbind(SS,df,MS)
```

Calculations by hand in R

```
##             SS df       MS
## [1,]  799.1381  1 799.1381
## [2,]  215.6119  2 107.8059
## [3,] 1014.7500  3 338.2500
```

Check calculations using builtin lm, summary and ANOVA commands in R

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##       253.4       -293.5
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      1      2      3      4
## -5.563  4.113  9.854 -8.404
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   253.43      10.77  23.537   0.0018 **
## x            -293.53     107.81  -2.723   0.1126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.38 on 2 degrees of freedom
## Multiple R-squared:  0.7875,Adjusted R-squared:  0.6813
## F-statistic: 7.413 on 1 and 2 DF,  p-value: 0.1126
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## x          1 799.14  799.14  7.4127 0.1126
## Residuals  2 215.61  107.81
##        1        2        3        4
## 247.5633 232.8868 221.1456 209.4043
##         1         2         3         4
## -5.563342  4.113208  9.854447 -8.404313
## [1] 215.6119
## [1]  799.1381  215.6119 1014.7500
```

Statistical inferences - CI's and tests for the β's

2. Standard Errors for Regression Coefficients

   (a) Regression coefficient values, $\hat{\beta}_0$ and $\hat{\beta}_1$, are point estimates of the true intercept and slope, β_0 and β_1 respectively.

   (b) To develop interval estimates (confidence intervals) for β_0 and β_1, we need to make assumptions about the errors in the regression model. In partiular, we assume $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ i.i.d $N(0, \sigma^2)$, in which case:

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{SS_{xx}})$$

   (c) The standard deviation of $\hat{\beta}_1$ is $\sqrt{\frac{\sigma^2}{SS_{xx}}}$

   (d) The value of σ^2 is unknown, so the estimator MSE is used in its place to produce the standard error of the estimate $\hat{\beta}_1$, as

$$SE_{\hat{\beta}_1} = \sqrt{MSE/SS_{xx}}$$

   (e) The standard error for estimate $\hat{\beta}_0$ is given as:

$$SE_{\hat{\beta}_0} = \sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}})}$$

   (f)
   - Standard Errors for regression coefficients in the above example are estimated below.
   - $SS_{xx} = .009275$ and $MSE = 107.80$
   - $SE_{\hat{\beta}_1} = \sqrt{MSE/SS_{xx}} = \sqrt{107.80/.009275} = 107.81$
   - $SE_{\hat{\beta}_0} = \sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}})} = \sqrt{107.80((1/4) + (.0399/.009275))} = 10.77$

3. Confidence Intervals for Regression Coefficients

   (a) Confidence intervals are constructed using the standard errors as follows:

$$\hat{\beta}_i \pm t_{\alpha/2,n-2} SE_{\hat{\beta}_i}$$

   (b) In the example, 95% confidence intervals for β_1 and β_0 are computed as follows.

   - $t_{\alpha/2,n-2} = t_{.025,2} = 4.303$
   - For the slope, β_1: $-293.531 \pm 4.303(107.81)$

$$(-757.4,\ \ 170.3)$$

   - For the intercept, β_0: $253.434 \pm 4.303(10.77)$

$$(207.1,\ \ 299.8)$$

95% Confidence intervals in R

- upper 2.5th percentile of t-dist'n with n-2 d.f.

```
MSE=SSE/(n-2)
t=qt(.975,n-2)  #upper .025'th percentile of t with n-2 df.
t
```

```
## [1] 4.302653
```

- 95%confidence interval for β_1

```
SEb1=sqrt(MSE/SXX) #standard error of beta_1
c(b1-t*SEb1,b1+t*SEb1)
```

```
## [1] -757.4057  170.3437
```

Why does the confidence interval have the correct coverage probability?

Consider the example of the interval for $\hat{\beta}_1$. We need the following facts:

(a) $\widehat{\beta}_1$ has a normal distribution with mean β_1 and unknown variance σ^2/SXX. A consequence is that $Z = \frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{SXX}} \sim N(0,1)$ (Easy results to prove.)

(b) $W = \frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$, a chi-squared distribution with $n-2$ degrees of freedom. (A bit harder to prove.)

(c) $\widehat{\beta}_1$ and SSE are independent, implies $Z = \frac{\widehat{\beta}_1 - \beta_1}{\sigma/\sqrt{SXX}}$ and $\frac{(n-2)MSE}{\sigma^2}$ are independent. (Hard to prove. Details involve considerable matrix algebra, and are contained in appendix C3 of Montgomery *et al*)

(d) Definition: If Z is standard normal, independent of W which is χ^2_ν, the $t = \frac{Z}{\sqrt{W/\nu}}$ is defined to have a t distribution with ν degrees of freedom.

(e) Then see general notes on constructing confidence intervals.

4. The **correlation** between X and Y is estimated by:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

An alternative expression is given by

$$r = \hat{\beta}_1 \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

or

$$r = \hat{\beta}_1 \frac{\sqrt{SS_{xx}}}{\sqrt{SS_{yy}}}$$

where $SS_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $SS_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ are the sums of squares of the X's and Y's, respectively. Note that $SS_{yy} = SST$, the total sum of squares. Note that $\frac{\sqrt{SS_{xx}}}{\sqrt{SS_{yy}}} = \frac{s_x}{s_y}$, the ratio of the standard deviations of the X's and the Y's.

- The correlation coefficient lies in the interval [-1,+1].
- If the relationship beween Y and X is perfectly linear and increasing, the correlation will be +1.
- If the relationship is perfectly linear and decreasing, the correlation will be +1. If there is no
- linear relationship between X and Y, the correlation is 0.
- In the example, $r = \hat{\beta}_1 \frac{\sqrt{SS_{xx}}}{\sqrt{SS_{yy}}} = -293.531 \frac{\sqrt{.009275}}{\sqrt{1016.49}} = -.887$

5. Goodness of fit of the regression line is measured by the **coefficient of determination**, R^2. For simple linear regression $R^2 = r^2$.

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

The Regression Sum of Squares (SSR) is similar to the Treatment Sum of Squares in an ANOVA problem. It is given by $SSR = \frac{SS_{xy}^2}{SS_{xx}}$. Alternative ways of calculating the residual sum of squares are to use the additivity relationship (SSR + SSE = SST), or to use one of the following formulas.

$$\begin{gathered} R^2 = SSR/SST \\ 1 - R^2 = (SST - SSR)/SST = SSE/SST \\ SSE = (1 - R^2)SST \end{gathered}$$

R^2 is the fraction of the total variability in y accounted for by the **linear** regression line, and ranges between 0 and 1. $R^2 = 1.00$ indicates a perfect linear fit, while $R^2 = 0.00$ is a complete linear non-fit.

In the example:

- $SSR = \frac{SS_{xy}^2}{SS_{xx}} = (-2.7255)^2/.009275 = 800.90$
- $SST = SSR + SSE = 800.90 + 215.59 = 1016.49$
- $R^2 = SSR/SST = 0.786$
- Note that $R^2 = r^2$, the square of the correlation coefficient.
- 78.8% of the variability in Y is accounted for by the regression model.

```
## [1] 799.1381
## [1] -0.8874245
## [1] 0.7875222
```

6. **Estimating the mean of Y**

   (a) The estimated mean of Y when $x = x^*$ is $\hat{\mu}_{x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^*$.

   (b)
   $$\hat{\mu}_{x^*} = \hat{\beta}_0 + \hat{\beta}_1 x^* \approx N\left(\beta_0 + \beta_1 x^*, \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right)\right)$$

   (c) The standard error of $\hat{\mu}_{x^*}$ is
   $$SE_{\hat{\mu}_{x^*}} = \sqrt{MSE\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right)}$$

   (d) A confidence interval for the mean $\mu_{x^*} = \beta_0 + \beta_1 x^*$ when $x = x^*$ is given by
   $$\hat{\mu}_{x^*} \pm t_{\alpha/2,n-2} SE_{\hat{\mu}_{x^*}}$$

   (e) eg. A 95% confidence interval for the mean at $x = 0.10$ is:

   - When $x^* = 0.10$, the estimated mean is $\hat{\mu}_{.1} = 253.434 - 293.531(0.1) = 224.08$
   - $SE_{\hat{\mu}_{.1}} = \sqrt{107.8\left(\frac{1}{4} + \frac{(0.1-.0875)^2}{.009275}\right)} = 5.36$
   - $t_{\alpha/2,n-2} = t_{.025,2} = 4.303$
   - margin of error $= 4.303(5.36) = 23.08$
   - 224.08 ± 23.08
   - (201 , 247.16)

95% confidence interval for mu at x0=.10

```
x0=.10
muhat=b0+b1*x0 # estimate of mean at x=x0
muhat
SEmu=sqrt(MSE)*sqrt(1/n+(x0-mean(x))^2/SXX) #SE of muhat
SEmu
c(muhat-t*SEmu, muhat+t*SEmu)
```

```
## [1] 224.0809
## [1] 5.363545
## [1] 201.0034 247.1583
```

7. **Predicting a New Response Value**

We are now interesting in predicting the value of y at a future value $x = x^*$. In making a **prediction interval** for a future observation on y when $x = x^*$, we need to incorporate two sources of variation which account for the fact that we are replacing the unknown mean by the estimate $\hat{\beta}_0 + \hat{\beta}_1 x^*$, and we are replacing the unknown standard deviation σ by the estimate $\sqrt{MSE}$.

$$y - (\hat{\beta}_0 + \hat{\beta}_1 x^*) = (y - (\beta_0 + \beta_1 x^*)) - (\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*))$$

The first term in brackets on the right hand side of this expression has a $N(0, \sigma^2)$ distribution. From (b) above, the distribution of the second term is

$$N\left(0, \sigma^2\left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right)\right)$$

As y represents a future observation, the distributions of the two terms are independent, and it follows that the distribution of $y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$ is

$$N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right)\right)$$

(a) The predicted value of y is given by $\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$

(b) The variance of the above distribution is estimated by:

$$\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right)}$$

(c) and the prediction interval for y is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2}\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SS_{xx}}\right)}$$

(d) eg. A 95% prediction interval for y when $x = 0.10$ is:

- For $x^* = 0.10$, $y^* = 253.434 - 293.531(0.1) = 224.08$
- $SE_{y^*} = \sqrt{107.8\left(1 + \frac{1}{4} + \frac{(0.1 - .0875)^2}{.009275}\right)} = 11.69$
- $t_{\alpha/2, n-2} = t_{.025,2} = 4.303$
- margin of error $= 4.303(11.69) = 50.29$
- 224.08 ± 50.29
- (173.79, 274.37)

```
SEmu=sqrt(MSE)*sqrt(1+1/n+(x0-mean(x))^2/SXX)
c(muhat-t*SEmu, muhat+t*SEmu)
```

95% prediction interval for a new observation at x0=.10

```
## [1] 173.7980 274.3637
```