Linear Regression

1. Least squares regression line

The statistical model for simple linear regression states that the response $y$ when the explanatory variable takes the value $x$ is

$$y = \beta_0 + \beta_1 x + \epsilon$$

Here $\beta_0 + \beta_1 x$ is the mean response. The random error $\epsilon$’s are assumed to be independent and normally distributed with mean 0 and standard deviation $\sigma$. The parameters of the model are $\beta_0$, $\beta_1$, and $\sigma$.

Given $n$ observations on the explanatory variables $x$ and the response variable $y$,

$$(x_1, y_1), \ldots, (x_n, y_n)$$

The least-squares regression line of $y$ on $x$ is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible. For any pair of $(a, b)$ as the estimate for $(\beta_0, \beta_1)$, we define the residuals as

$$e_i = \text{observed } y_i - \text{predicted } y_i = y_i - \hat{y}_i = y_i - a - bx_i.$$  

Suppose $(\hat{\beta}_0, \hat{\beta}_1)$ is the value that minimize $\sum e_i^2$. The least-squares line is: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ with slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = r \frac{s_y}{s_x}$$

and intercept

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

Here $\bar{x}$ and $s_x$ are the sample mean and sample standard deviation of $X$.

$$s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}.$$  

Similarly, $\bar{y}$ and $s_y$ are the sample mean and sample standard deviation of $Y$. Meanwhile, $r$ is the Pearson correlation coefficient defined as

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right).$$

How about $\sigma$?

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum(y_i - \hat{y}_i)^2}{n-2}.$$  

Then, $s^2$ is an unbiased estimate of $\sigma^2$. $n - 2$ is called the degree of freedom for $s^2$. The estimate of $\sigma$ is $s = \sqrt{s^2}$. 
2. Example

We have a regression of \( Y = \text{reaction time} \) on \( X = \text{percent of drug} \), where \( X = (1, 2, 3, 4, 5) \) and \( Y = (1, 1, 2, 2, 4) \). It is easy to compute that

\[
\bar{x} = 3, \ \bar{y} = 2, \ SS_{xx} = \sum (x_i - \bar{x})^2 = 10, \ SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = 7.
\]

Thus, \( \hat{\beta}_1 = SS_{xy}/SS_{xx} = 0.7 \), and \( \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = -0.1 \). Therefore, the \( SSE = \sum (y_i - \hat{y}_i)^2 = 1.1 \), \( s^2 = SSE/(n - 2) = 1.1/3 = 0.3667 \), and \( s = \sqrt{0.3667} = 0.6056 \).

If we are interested in construct 95% confidence interval for \( \beta_1 \), we first find the standard error of \( \hat{\beta}_1 \), i.e. \( se(\hat{\beta}_1) = s/\sqrt{SS_{xx}} \). we have

\[
\hat{\beta}_1 \pm t_{n-2, \alpha/2} \cdot s/\sqrt{SS_{xx}} \leftrightarrow 0.7 \pm 3.1824 \cdot 0.6056/\sqrt{10} 
\]

where 3.0824 is the 97.5% quantile of a t-distribution with \( df = 3 \). The hypothesis testing associated with \( \beta_1 \) proceeds in a similar way.

Conditioning on \( X = x^* \), we can also make prediction for the conditional mean \( E(Y|X = x^*) \) or the individual observation \( Y|X = x^* \). See text book 11.6 for details. Given \( x = x^* \), the predicted conditional mean is

\[
\hat{E}(Y|X = x^*) = \hat{\beta}_0 + \hat{\beta}_1 x^*
\]

with

\[
se(\hat{E}(Y|X = x^*)) = s \cdot \sqrt{1/n + (x^* - \bar{x})^2/SS_{xx}}.
\]

The predicted value for the individual observation is

\[
\hat{Y}|X = x^* = \hat{\beta}_0 + \hat{\beta}_1 x^*
\]

with

\[
se(\hat{Y}|X = x^*) = s \cdot \sqrt{1 + 1/n + (x^* - \bar{x})^2/SS_{xx}}.
\]

For a linear regression problem, we can also put it under the framework of ANOVA, where a F-test can be used to assess the utility of the model.

Data = Fit + Residual

\[
y_i = \hat{y}_i + (y_i - \hat{y}_i)
\]

Therefore,

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

Algebraic fact: \( \sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \).

Rewrite this as \( SST = SSM + SSE \).

Degree of Freedoms: \( DFT = DFM + DFE \)
where \( DFT = n - 1 \), \( DFM = 1 \), and \( DFE = n - 2 \).

**Mean Square**  
\[ MS = \frac{\text{sum of squares}}{\text{degrees of freedom}} \]  
for example,

\[ MSE = s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2} \]

The **coefficient of determination** \( r^2 \) is the percentage of total variability can be explained by the model.

\[ r^2 = \frac{SSM}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} \]

The overall **ANOVA F-Test** \( F = \frac{MSM}{MSE} \) compared with a F distribution with \( df = (1, n-2) \), where the null is \( H_0 : \beta_1 = 0 \) vs. \( H_a : \beta_1 \neq 0 \). Recall a \( F_{df=(1,k)} \) is the square of a \( t_{df=k} \).

If use software R, here is the code and printout:

```r
> x<- c(1,2,3,4,5)
> y<- c(1,1,2,2,4)
> fit1<- lm(y~x)
> summary(fit1)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.1000 0.6351 -0.157 0.8849
x 0.7000 0.1915 3.656 0.0354

Residual standard error: 0.6055 on 3 degrees of freedom
Multiple R-squared: 0.8167, Adjusted R-squared: 0.7556
F-statistic: 13.36 on 1 and 3 DF, p-value: 0.03535
```

```r
> anova(fit1)

Analysis of Variance Table

Response: y
Df Sum Sq Mean Sq F value Pr(>F)
x 1 4.9000 4.9000 13.364 0.03535 *
Residuals 3 1.1000 0.3667

## construct 95% confidence interval for conditional mean and individual obs.
## at x=3.5.
> meanyhat<- predict(fit1, data.frame(x=c(3.5)),interval="confidence")
> obsyhat<- predict(fit1, data.frame(x=c(3.5)),interval="prediction")
> meanyhat

fit lwr upr
1 2.35 1.435912 3.264088
> obsyhat

fit lwr upr
1 2.35 0.2171277 4.482872

## you may create the confidence bands to compare them.
> plot(x,y, ylim=c(-1, 6))
> abline(-0.01, 0.70)
> pred.mean<- predict(fit1, interval="confidence")
```
2. Multiple linear regression

The MLR model is

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon, \]

where \( x_1, \ldots, x_k \) are \( k \) regressors/predictors. Like in SLR, \( \epsilon \)'s follow independent and identical normal distributions with mean 0 and standard deviation \( \sigma \). Follow the least squares criterion, the estimation and testing of coefficients \( \beta_1, \ldots, \beta_k \) are very much the same as those for simple linear regressions. But the degree of freedom of \( SSE \) becomes \( n - (k + 1) \). Therefore, for construction or testing associated with any \( \beta_j \), we would use the quantiles of \( t \)-distribution with \( df = n - (k + 1) \). Similarly, the overall F-test \( F = \frac{MSM}{MSE} \) compared with a \( F \) distribution with \( df = (k, n - (k + 1)) \), where the null is \( H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \) vs. \( H_a : \text{some } \beta_j \neq 0 \).

However, there is one feature of MLR is unique compared with SLR, which is model selection. Suppose we have two models one of which is a special case of the other. We call them nested models. For example,

Reduced model: \( E[y|X] = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g \);
Complete model: \( E[y|X] = \beta_0 + \beta_1 x_1 + \ldots + \beta_g x_g + \beta_{g+1} x_{g+1} + \ldots + \beta_k x_k \),

where \( g < k \). The reduced model is a spacial case of the complete model when \( \beta_{g+1} = \ldots = \beta_k = 0 \). Therefore, to compare these models is equivalent to test \( H_0 : \beta_{g+1} = \ldots = \beta_k = 0, vs H_a : \text{some } \beta_j \neq 0, g + 1 \leq j \leq k \). The test statistic:

\[ F = \frac{(SSE_{\text{reduced}} - SSE_{\text{complete}})/(k - g)}{SSE_{\text{complete}}/[n - (k + 1)]}, \]

where \( F \) should be compared with a \( F \)-distribution with \( df = (k - g, n - k - 1) \). Suppose we set \( \alpha = 0.05 \). If the test statistic exceeds the 95\% quantile of \( F_{(k-g,n-k-1)} \) distribution, then we reject the null and prefer the complete model over the reduced one. The consecutive subscripts in the models are used only for convenience. The same logic and test also apply on comparison of the reduced model \( E[y|X] = \beta_0 + \beta_2 x_2 + \beta_4 x_4 \) and the complete model \( E[y|X] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_5 x_5 \).

MLR can also be used for qualitative/categorical predictors. But we have to code the values of the qualitative predictor. These coded qualitative variables are called dummy/indicator variables. For example, in one way ANOVA discussed before, we have one treatment have \( k \) levels. We can create \( k - 1 \) dummy variable: \( x_i = 1 \) if \( y \) is observed at level \( i + 1 \) otherwise \( x_i = 0 \), for \( i = 1, 2, \ldots, k - 1 \). A MLR of \( y \) on these \( k - 1 \) dummy variables exactly reproduce the one way ANOVA.