Simple Linear Regression

Regression	Studies the relationship between quantitative variables.
Simple Linear Regression	Only considers 2 variables
Response Variable	Usually denoted Y. We attempt to predict this.
Predictor Variable	Usually denoted X. We use this to predict Y.
(x_i,y_i)	The values for X and Y at case i. We usually denote n to be the number of cases.
Outline of Simple Linear Regression	Assume a linear relationship between X and Y: $Y = \beta_0+\beta_1$
β_0	The intercept ie. the value of Y when X=0, ie where the line crosses the Y axis.
β_1	The slope. The change in Y for a single unit change in X.
	We estimate β_0 and β_1 from the data and use the model to predict Y for any given X.
Methods of Linear Regression	
Scatter Plot	Put all points on a scatter plot and gauge visually whether or not the relationship looks linear.
Line of Closest Fit	If the relationship looks linear then we find the line fo closest fit and use it to estimate β_0 and β_1

Co-Variance and Independent Variables

Independent Events	P(A|B)=P(A)
Independent Discrete Variables	P(X = x and Y=y) = P(X=x)P(Y=y)
Independent Continuous Variables	The joint pdf of X and Y = h(x,y) = fx(x)gy(y) - the product of individual pdfs.
Covariance	"the mean value of the product of the deviations of two variates from their respective means" Covariance of X and Y = cov(X,Y) = $E(X - \mu_1)(Y - \mu_2)$ where $\mu_1=E(X)$ and $\mu_2=E(Y)$

Co-Variance and Independent Variables (cont)

Covariance of independent variables	cov(X,Y)=0
Covariance as defined by the book	Measures the association between X and Y, the extent to which they vary together. If large X occurs with large Y and small x with small y, there is a positive association ie. cov(X,Y) > 0. If large X occurs with small y and Large Y occurs with small x, there is a negative associationie. cov(X,Y) < 0.
Direction of association	+ indicates postive direction, - indicates negative direction.

Least Squares Criterion

Intro	In a scatter plot there could be many potential lines that could fit the data. We use the Least SquaresCriterion to select the best line.
e_i (error)	THe difference between what the line says the value should be and what it actually is.
e_i (residual)	Difference between the fitted line and actual reality
Residual Sum of Squares (RSS)	We chose β_0 and β_1 so as to minimize RSS.

^ above a letter indicates we are using an estimator

Least Sum of Squares Important Formula

$$RSS = min(\sum_{i=1}^{n} \hat{e}_i^2)$$

$$Through\ Partial\ differentiation\ we\ derive\ the\ estimators...$$

$$\hat{\beta}_1 = \frac{SXY}{SXX}$$

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

RSS=min(\sum_{i=1}^n\hat{e}_i^2)\\Through\ Partial\ differentiation\ we\ derive\ the\ estimators...\\\hat{β}_{1}=\frac{SXY}{SXX}\\\hat{β}_{0}=\bar{y}-\hat{β}_{1}\bar{x}

Errors

Real data almost never falls in a perfectly straight line. ie. Real data rarely has a perfectly linear relationship. As such real data has errors which could be...
- Measurement Errors: Continuous Variables cannot be measured with 100% accuracy.
- An effect of variables not included in the model
- Natural variability.

We should incorporate them into our simple linear regression models. eg.
$y_i = \beta_0 + \beta_1 x_i + e_i$ where e_i is the error on the ith case
and
$y_i = \beta_0 + \beta_1 x_i$ is the true regression line
*
Assumptions about errors:
We make these assumptions as we need them to...
- prove the optimaity ofthe estimates for β_0 and β_1
- prove the confidence intervals for β_0 and β_1

$e_i \sim NID(0,\sigma^2)$
- N: Normally distributed with mean 0
- I: Independent variables
- D: Distributed.
- σ^2: Common Variance.
- "e_i is normally distributed with mean 0 and common variance of σ^2"

These assumption can also be expressed in terms of "Co-Variance"
$E(e_i) = 0$, $var(e_i) = \sigma^2$, $cov(e_i,e_j) = 0$, for $i \neq j$
- "Expected value e_i is 0, variance is σ^2, covariance of e_i and e_j is 0 where i is not j"

Combined with the normality assumptions, this implies e_is are independent.
Assumptions must be verified when applying to a regression model.

Sample Correlation Coefficient rxy

r_{xy} =	SXY/sqrt((SXX)(SYY)) = [SXY/(n-1)/[sqrt((SXX/n-1)(SYY/n-1))]
Correlation Coefficient	r_{xy} is the sample covariance scaled to lie in[-1,1]. ie. -1<=r_{xy}<=1
$r_{xy}>0$	Positive association
$r_{xy}<0$	Negative association
$r_{xy}=1$	All points lie on positive slope. The closer r_{xy} is to 1, the closer all points are to lying on the positive line.
$r_{xy}=-1$	All points lie on negative slope. The closer r_{xy} is to -1, the closer all points are to lying on the negative line.
Bivariate Regression	rr and/or its square r^2 is used to measure howwell the linear model fits the data.

Sample Correlation Coefficient rxy (cont)

Multiple Regression	The multiple correlation coefficient (R^2) is used to measure how well the linear model fits the data.
x-bar/$\bar{x}$	Indicates the sample mean of x
SXY	The standard deviation of X on Y
Linearity	Linearity cannot be deduuced from correlation coefficient. It should be paired with the scatter plot and never be considered in isolation.

The X2 Distribution (Chi-Squared)

Degrees of Freedom (df)	The number of different values/quantities which a distribution can be assigned.
$X^2(v)$	A chi-squared distribution with v df.
$E(X^2(v)) = v$	ie. The expeced value of a X^2 distribution with v df, is v.
$RSS/\sigma^2 \sim X^2(n-2)$	So... $E(RSS/\sigma^2) => E(X^2(n-2)) = n-2$ and so $E(RSS/n-2)=\sigma^2$
RSS/n-2	An unbiased estimate of σ^2.
$sqrt(\sigma^2) = \sigma$	Estimate of Standard Error of Regression/Residual Standard Error(in R)
sqrt(estimated variance) =	standard error