Regression Analysis

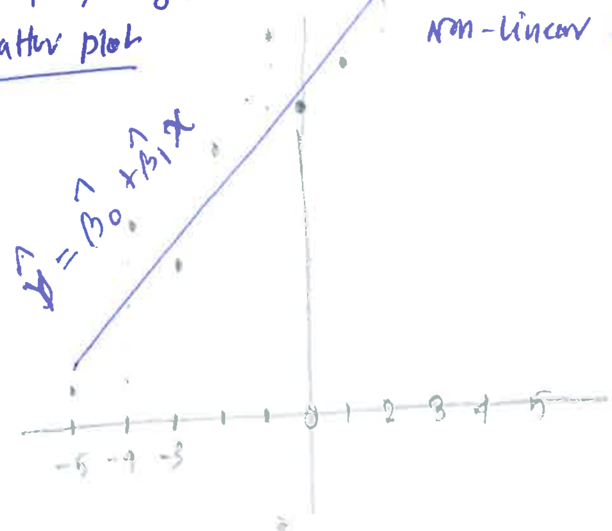
1

Ref: (1) Draper & Smith, Applied Regression Analysis

(2) Montgomery, Peck, Vining Introduction to Linear Regression Analysis.

- Simple linear Regression
- Multiple linear regression
- Selecting the best regression model
- Multicollinearity
- Model Adequacy checking
- Test for influential observations.
- Transformation & weighting to correct model inadequacies
- Dummy variables
- poly reg. models; Generalized linear model
- Non-linear estimation.

Scatter plot



Simple linear regression

Simple linear regression model is $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

β_0 intercept $\quad \epsilon_i$ error.

β_1 slope

Regression Analysis is a statistical tool for investigating the relationship between a dependent variable and one or more independent variables.

Temperature (X)	Yield (Y)
−5	1
−4	5
−3	4
−2	7
−1	10
0	8
1	9
2	13
3	14
4	13
5	18
Regressor variable / independent	Response variable / dependent

Basic Assumptions on the model

(1) ϵ_i is random variable with $E(\epsilon_i)=0$ & $V(\epsilon_i)=\sigma^2$ (unknown)

2. ϵ_i & ϵ_j are uncorrelated $Cov(\epsilon_i, \epsilon_j)=0$

3. ϵ_i are normally distributed r.v.s

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

$E(Y_i) = \beta_0 + \beta_1 x_i$

$V(Y_i) = \sigma^2$

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

Least Squares estimation of the parameters:

We estimate β_0 & β_1 so that Sum of squares of the diff. between y_i & straight line is min.

$$S = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$= \sum_{i=1}^{n} e_i^2 = SS_{Res}$$

$$\sum x_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = 0$$

$$\sum x_i (y_i - \bar{y} - \hat{\beta}_1 (x_i - \bar{x})) = 0$$

$$\hat{\beta}_1 = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})}$$

$$= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$= \frac{S_{xy}}{S_{xx}}$$

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = 0 \Rightarrow$$

$$\sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = 0 \Rightarrow$$

$$\sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\hat{\beta}_0 = \frac{1}{n}\sum y_i - \hat{\beta}_1 \frac{\sum x_i}{n}$$

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

Normal equations

(1) $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$

$\Rightarrow \sum_{1}^{n} e_i = 0$ $\boxed{\sum y_i = \sum \hat{y}_i}$

(2) $\sum_{i=1}^{n} e_i x_i = 0$

(3) $\sum e_i \hat{y}_i = 0$

$$\sum e_i (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= \hat{\beta}_0 \sum e_i + \hat{\beta}_1 \sum e_i x_i$$

$$= 0$$

But $\sum e_i y_i \neq 0$

Statistical properties of $\hat{\beta}_0$ & $\hat{\beta}_1$

$\hat{\beta}_0$ & $\hat{\beta}_1$ are linear combination of observations y_i.

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

$$= \frac{\sum y_i (x_i-\bar{x})}{\sum (x_i-\bar{x})^2}$$

$$= \sum y_i c_i \quad \text{where}$$

$$c_i = \frac{x_i-\bar{x}}{\sum (x_i-\bar{x})^2}$$

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

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

$$E(\hat{\beta}_0) = E(\beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} - \hat{\beta}_1 \bar{x})$$

$$= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} = \beta_0$$

$$V(\hat{\beta}_1) = V\left(\frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}\right)$$

$$= V\left(\frac{\sum (x_i-\bar{x}) y_i}{\sum (x_i-\bar{x})^2}\right)$$

$$= V\left(\sum c_i y_i\right)$$

where $c_i = \dfrac{(x_i-\bar{x})}{\sum (x_i-\bar{x})^2}$

③ $\hat{\beta}_0$ & $\hat{\beta}_1$ are unbiased estimator of β_0 & β_1 resp.

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

$$y_i - \bar{y} = \beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \bar{x} - \bar{\varepsilon} + \varepsilon_i$$

$$= \beta_1 (x_i-\bar{x}) + (\varepsilon_i - \bar{\varepsilon})$$

$$E(y_i-\bar{y}) = \beta_1 (x_i-\bar{x})$$

$$E(\hat{\beta}_1) = \frac{\beta_1 \sum (x_i-\bar{x})^2}{\sum (x_i-\bar{x})^2} = \beta_1$$

$$= \frac{\sum (x_i-\bar{x}) E(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$$

$$V\left(\sum c_i y_i\right) = \sum c_i^2 \sigma^2$$

$$= \frac{\sigma^2 \sum (x_i-\bar{x})^2}{\left(\sum (x_i-\bar{x})^2\right)^2}$$

$$= \frac{\sigma^2}{\sum (x_i-\bar{x})^2}$$

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

$$= V(\bar{y}) + \bar{x}^2 V(\hat{\beta}_1) - 2\,\mathrm{cov}(\bar{y}, \hat{\beta}_1 \bar{x})$$

$$= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{S_{xx}}$$

(4)

$$\operatorname{Cov}(\bar{y}, \hat{\beta}_1)$$

$$= \operatorname{Cov}\left(\frac{\sum y_i}{n}, \frac{\sum (x_i-\bar{x}) Y_i}{\sum (x_i-\bar{x})^2}\right)$$

$$= \frac{\sum (x_i-\bar{x}) V(Y_i)}{n \sum (x_i-\bar{x})^2}$$

$$= \frac{\sigma^2 \sum (x_i-\bar{x})}{n \sum (x_i-\bar{x})^2} = 0$$

Estimation of σ^2

$$SS_{Res} = \sum e_i^2 = \sum (y_i-\hat{y}_i)^2$$

$$= \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$= \sum (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)^2 \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

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

$$= \sum (Y_i-\bar{Y})^2 + \hat{\beta}_1^2 \sum (x_i-\bar{x})^2 - 2\hat{\beta}_1 \sum (Y_i-\bar{Y})(x_i-\bar{x})$$

$$= S_{YY} + \hat{\beta}_1^2 S_{XX} - 2\hat{\beta}_1 S_{XY}$$

$$= S_{YY} + \hat{\beta}_1^2 S_{XX} - 2\hat{\beta}_1^2 S_{XX} \qquad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

$$= S_{YY} - \hat{\beta}_1^2 S_{XX}$$

$$E(S_{YY}) = E\sum (Y_i-\bar{Y})^2$$

$$= E\left(\sum Y_i^2 - n\bar{Y}^2\right)$$

$$= \sum E(Y_i^2) - n E(\bar{Y}^2)$$

$$E(Y_i^2) = V(Y_i) + E^2(Y_i)$$

$$= \sigma^2 + (\beta_0 + \beta_1 x_i)^2$$

$$E(\bar{Y}^2) = V(\bar{y}) + (E(\bar{Y}))^2$$

$$= \frac{\sigma^2}{n} + (\beta_0 + \beta_1 \bar{x})^2$$

$$= n\sigma^2 + \sum (\beta_0 + \beta_1 x_i)^2 - \sigma^2 - n(\beta_0 + \beta_1 \bar{x})^2$$

$$= (n-1)\sigma^2 + \beta_1^2 \sum (x_i-\bar{x})^2$$

$$= (n-1)\sigma^2 + \beta_1^2 S_{xx}$$

$$E(\hat{\beta}_1^2 S_{XX}) = S_{XX} \cdot E(\hat{\beta}_1^2)$$

$$= S_{XX}\left[V(\hat{\beta}_1) + (E(\hat{\beta}_1))^2\right]$$

$$= S_{XX}\left[\frac{\sigma^2}{S_{XX}} + \beta_1^2\right]$$

$$= \sigma^2 + \beta_1^2 S_{XX}$$

$$E(SS_{Res}) = E(S_{YY}) - E(\hat{\beta}_1^2 S_{xx})$$

$$= (n-1)\sigma^2 + \beta_1^2 S_{xx} - \sigma^2 - \beta_1^2 S_{xx}$$

$$= (n-2)\sigma^2$$

$$E\left(\frac{SS_{Res}}{n-2}\right) = \sigma^2$$

$$\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$$

Test Statistic

$$Z = \frac{\hat{\beta}_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \quad \text{under } H_0: \beta_1 = 0$$

If σ^2 is known, we can use Z to test the hypothesis $H_0: \beta_1 = 0$

Reject H_0 if

$$|Z| > Z_{\alpha/2}$$

Test of Slope coefficient

$H_0: \beta_1 = 0$ (no linear relationship)

$H_1: \beta_1 \neq 0$ (linear relationship)

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \sum c_i Y_i$$

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

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

Then $$Z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

Usually σ^2 is not known

$$\sigma^2 = E(MS_{Res}) = E\left(\frac{SS_{Res}}{n-2}\right)$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}}$$

$$= \frac{\hat{\beta}_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2} \quad \text{under } H_0: \beta_1 = 0$$

We reject H_0: if $|t| > t_{\alpha/2, n-2}$

$$\left.\begin{array}{l} \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1) \\[2ex] \dfrac{(n-2) MS_{Res}}{\sigma^2} \sim \chi^2_{n-2} \end{array}\right\} \text{ind.}$$

The Analysis of Variance

ANOVA

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

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

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

$$= \sum \hat{y}_i e_i - \bar{y}\sum e_i$$

$$= 0-0=0$$

$$\underbrace{\sum(y_i-\bar{y})^2}_{\text{Total variation in the observations}} = \underbrace{\sum(\hat{y}_i-\bar{y})^2}_{\text{Regression Sum of Squares}} + \underbrace{\sum(y_i-\hat{y}_i)^2}_{\text{Residual Sum of Squares}}$$

$$SS_{Res} \sim \chi^2_{n-2}$$

SS_T has degree of freedom $n-1$

$$SS_T = \sum_{i=1}^{n}(Y_i-\bar{Y})^2$$

$(Y_1-\bar{Y})$

$(Y_2-\bar{Y})$

$(Y_n-\bar{Y})$ Sanity

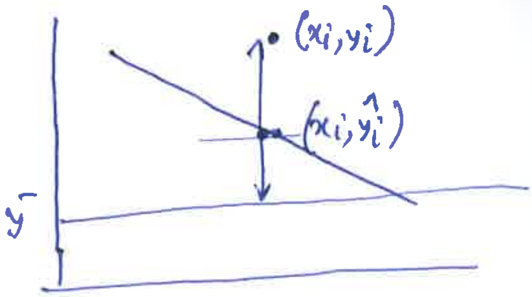
$$\sum(Y_i-\bar{Y})=0$$

$$SS_T = SS_{Reg} + SS_{Res}$$

$$SS_{Res} = S_{YY} - \hat{\beta}_1^2 S_{XX}$$

$$= SS_T - SS_{Reg}$$

$$\boxed{SS_{Reg} = \hat{\beta}_1^2 S_{XX} = \hat{\beta}_1 S_{XY}}$$



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

$$SS_T = SS_{Reg} + SS_{Res}$$

$$DF_T = DF_{Reg} + DF_{Res}$$

$$n-1 = 1 + n-2$$

ANOVA Table

Source of variation	DF	SS	MS	F
Reg	1	SS_{Reg}	MS_{Reg}	$F=\frac{MS_{Reg}}{MS_{Res}}$
Residual	$n-2$	SS_{Res}	MS_{Res}	
Total	$n-1$	SS_T		

$E(MS_{Res}) = \sigma^2$

$E(MS_{Reg}) = \sigma^2 + \beta_1^2 S_{xx}$

$$\left.\begin{aligned} &\frac{(n-2)(MS_{Res})}{\sigma^2} \sim \chi^2_{n-2} \\ &\frac{MS_{Reg}}{\sigma^2} \sim \chi^2_1 \quad \text{under } H_0: \beta_1 = 0 \end{aligned}\right\} \text{ind.}$$

(M) Let $\left.\begin{aligned} X &\sim \chi^2_m \\ Y &\sim \chi^2_n \end{aligned}\right\}$ ind.

then $F = \dfrac{X/m}{Y/n} \sim F_{m,n}$

Case: $R^2 = 0$ when $SS_T = SS_{Res}$

$$\sum (Y_i - \bar{Y})^2 = \sum (Y_i - \hat{Y}_i)^2$$

$$\Leftrightarrow \hat{Y}_i = \bar{Y}$$

there is no relationship between Y and X.

$$R^2 = \frac{SS_{Reg}}{SS_T} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = 0$$

$$\Rightarrow \hat{\beta}_1 = 0.$$

$F = \dfrac{MS_{Reg}}{MS_{Res}} \sim F_{1,\, n-2}$

To test $H_0: \beta_1 = 0$ we compute F & reject H_0 if

$$F > F_{\alpha, 1, n-2}.$$

Coefficient of Determination

$R^2 = \dfrac{SS_{Reg}}{SS_T}$ = proportion of variability in response variance that is explained by the model.

$$0 \le R^2 \le 1$$

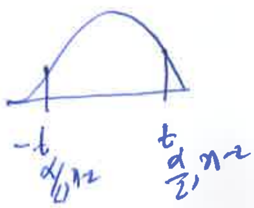
Confidence interval for β_1

LSE of β_1 is $\hat{\beta}_1 = \dfrac{S_{xY}}{S_{\cdot}}$

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

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \sim t_{n-2}$$



$$P\left\{ -t_{\alpha/2, n-2} < \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{MS_{Res}}{S_{xx}}}} \leq t_{\alpha/2, n-2} \right\} = 1-\alpha$$

$100(1-\alpha)\%$ CI for β_1 is

$$\hat{\beta}_1 - \sqrt{\frac{MS_{Res}}{S_{xx}}}\, t_{\alpha/2, n-2} < \beta_1 < \hat{\beta}_1 + \sqrt{\frac{MS_{Res}}{S_{xx}}}\, t_{\alpha/2, n-2}$$

Interval estimation on

Expected response $E(Y)$ for $X = X_0$.

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$E(Y) = \beta_0 + \beta_1 X$$

$$E(Y|X = X_0) = \beta_0 + \beta_1 X_0$$

$$\widehat{E(Y|X=X_0)} = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

is an unbiased estimator of $E(Y|X=X_0)$.

$$V(\hat{\beta}_0 + \hat{\beta}_1 X_0) \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$= V(\bar{y} + \hat{\beta}_1 (X_0 - \bar{x}))$$

$$= \frac{\sigma^2}{n} + \frac{(X_0 - \bar{x})^2 \sigma^2}{S_{xx}} + \underbrace{2\,\mathrm{Cov}(\bar{y}, \hat{\beta}_1)(X_0 - \bar{x})}_{0}$$

$$= \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{x})^2}{S_{xx}} \right)$$

$$\widehat{E(Y|X=X_0)} \sim N\left(\beta_0 + \beta_1 X_0,\ \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{x})^2}{S_{xx}} \right) \right)$$

$$\frac{\widehat{E(Y|X=X_0)} - E(Y|X=X_0)}{\sqrt{MS_{Res} \left(\frac{1}{n} + \frac{(X_0 - \bar{x})^2}{S_{xx}} \right)}} \sim t_{n-2}$$

$100(1-\alpha)\%$ CI for $E(Y|X=X_0)$

is

$$\widehat{E(Y|X=X_0)} \pm t_{\alpha/2, n-2} \sqrt{MS_{Res} \left(\frac{1}{n} + \frac{(X_0 - \bar{x})^2}{S_{xx}} \right)}$$

8

Predicted response y_0 at $x = x_0$

$$y_0 = \beta_0 + \beta_1 x_0 + \varepsilon_0$$

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

Define $\psi = y_0 - \hat{y}_0$

$$E(\psi) = E\left(\beta_0 + \beta_1 x_0 + \varepsilon_0 - \hat{\beta}_0 - \hat{\beta}_1 x_0\right)$$

$$V(\psi) = V(y_0) + V(\hat{y}_0)$$

$$= \sigma^2 + V(\hat{\beta}_0 + \hat{\beta}_1 x_0)$$

$$= \sigma^2 + V\left(\bar{y} + \hat{\beta}_1 (x_0 - \bar{x})\right)$$

$$= \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{S_{xx}} (x_0 - \bar{x})^2$$

y_0 & $\hat{y}_0$ are independent

The r.v.

$$\psi = y_0 - \hat{y}_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$$

$$\frac{y_0 - \hat{y}_0}{\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$$

$100(1-\alpha)\%$ PI on a future observation at x_0 is

$$\hat{y}_0 \pm \sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$$

To test $\beta_0 = a$ ag. $\beta_0 \neq a$.

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

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

$$t = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$

~~Under H_0~~

Under H_0

$$t = \frac{\hat{\beta}_0 - a}{\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$

Reject H_0 if

$$|t| > t_{\alpha/2, n-2}$$

$100(1-\alpha)\%$ Confidence interval on the intercept β_0 is

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}, n-2} \sqrt{MS_{Res}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}$$