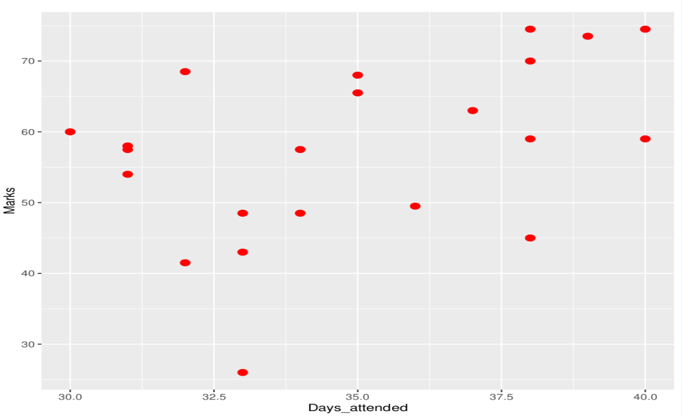
Statistical Inference (MA862)

Lecture Slides

Topic 5: Linear Regression

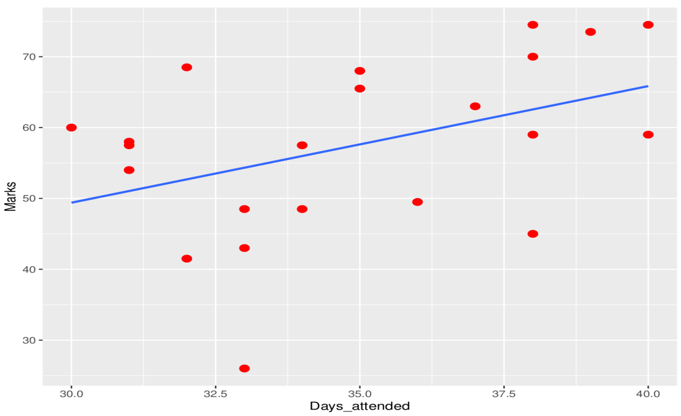
Regression

- Question: What is the impact of attending classes on students' final marks?
- Let's start with a real data from IITG which you can feel about it!!



Regression

- Question: What is the impact of attending classes on students' final marks?
- Let's start with a real data from IITG which you can feel about it!!



Linear Regressions

- We have one particular variable that we are interested in understanding or modeling, such as sales of a particular product, sale price of a home, or voting preference of a particular voter. This variable is called the target, response, or dependent variable, and is usually represented by y.
- We have a set of p other variables that we think might be useful in predicting or modeling the target variable (for *e.g.* the price of the product, the competitor's price, and so on; or the lot size, number of bedrooms, number of bathrooms of the home, and so on; or the gender, age, income, party membership of the voter, and so on). These are called the predicting, independent variables, or features and are usually represented by $x_1, x_2, \ldots, x_p$.

Linear Regressions

- Thus, we have

  $$y = f(\boldsymbol{x}; \boldsymbol{\beta}) + \varepsilon,$$

  for some real valued function f, where $\boldsymbol{x}$ is vector of predictors, $\boldsymbol{\beta}$ is the vector of parameters, and ε is error.
- If f is linear in the parameters vector $\boldsymbol{\beta}$, then the regression is called linear regression.

Examples and Role of Transformations

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$ is a linear model.
- $y = \beta_0 + \beta_1 x + \beta_2 x^2$ is a linear model, because it is linear in β (even though not in x).
- $y = \beta_0 + \beta_1 x^{\beta_2}$ is a non-linear model, as it is not linear in β.
- $y = \beta_0 x^{\beta_1}$ is not a linear model, but $\ln y = \ln \beta_0 + \beta_1 \ln x$ is.
- $y = \frac{e^{\beta x}}{1+e^{\beta x}}$, where $y \in (0,1)$
- $y = \frac{1}{\beta_0+\beta_1 x_1+\beta_2 x_2}$.

Main use of Linear Regressions

Typically, a regression analysis is used for one (or more) of three purposes:

1. modeling the relationship between x and y;
2. prediction of the target variable (forecasting);
3. testing of hypotheses.

Simple Linear Regression

- Just one predictor x, i.e. $p = 1$.
- The model for the simple linear regression is given by

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

  where y is the outcome variable (random), x is the independent/predictor variable (non-random) and ϵ is the random error term. β_0 (intercept) and β_1 (slope) are model parameters (unknown constants).

- Equivalently, the model can be written for $i = 1, 2, \cdots, n$ number of observations $(x_1, y_1), \cdots, (x_n, y_n)$ as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots, n.$$

- How do you interpret β_0 (intercept) and β_1 (slope)?

Least Squares Estimation

- Goal: To estimate β_0, β_1 by minimizing error in some sense (*e.g.* squared error)
- One reasonable way is to use the principle of Least Squares, *i.e.* minimize the objective function

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

  with respect to β_0, β_1.
- Differentiate $Q(\beta_0, \beta_1)$ with respect to β_0, β_1 and equate the partial derivatives to zero to get the estimates $\hat{\beta}_0, \hat{\beta}_1$.
- The resulting equations are called normal equations:

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \text{ and } \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Least Squares Estimation

- The solution is given by

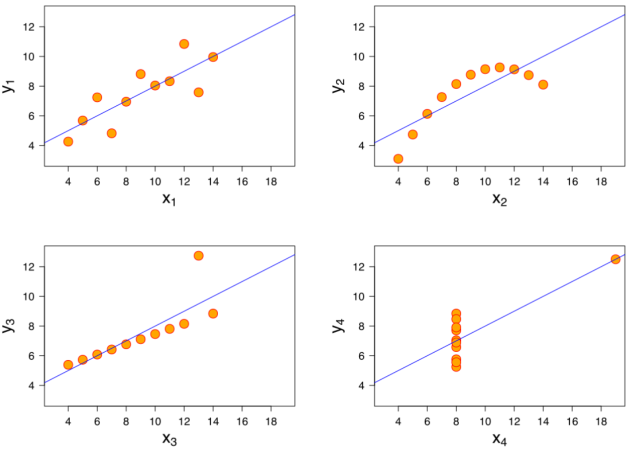
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{and} \quad \hat{\beta}_1 = S_{xy}/S_{xx}$$

where

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \;\text{ and }\; S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}).$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least squares estimator (LSE) of β_0 and β_1, respectively.

Importance of graphing data before analyzing it



Which one of the above do you think has highest value of absolute correlation and ideal for linear regression?

Importance of graphing data before analyzing it

- In all the four graphs: mean of x = 9 (with variance 11); mean of y = 7.50 (with variance 4.1) ; correlation between x and y = 0.816
- Fitted linear regression in each cases: $y = 3 + 0.5x$
- In 1973, Anscombe demonstrated the importance of graphing data before analyzing it and the effect of outliers on statistical properties

Assumptions

1. The errors are uncorrelated with each other, *i.e.*,

$$Cov(\epsilon_i, \epsilon_j) = 0 \text{ for } i \neq j.$$

2. The expected value of the errors is zero, *i.e.*,

$$E(\epsilon_i) = 0 \text{ for all } i.$$

3. The errors are homoscedastic (constant variance), *i.e.*,

$$Var(y_i) = \sigma^2 \text{ for all } i$$

.

Properties of $\hat{\beta}_0$ and $\hat{\beta}_1$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear combination of y_i's.
- $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased for β_0 and β_1, respectively.
- Variances of $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$
\begin{aligned}
Var(\hat{\beta}_0) &= \sigma^2\Big(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\Big) \\
Var(\hat{\beta}_1) &= \frac{\sigma^2}{S_{xx}}
\end{aligned}
$$

Gauss-Markov Theorem

Definition 5.1: An estimator $\hat{\theta}$ of θ is called linear estimator of θ if $\hat{\theta}$ is a linear combination of random observations.

Definition 5.2: An estimator $\hat{\theta}$ of θ is called the best linear unbiased estimator (BLUE) of θ if $\hat{\theta}$ is linear and unbiased estimator of θ and $\hat{\theta}$ has minimum variance among all linear unbiased estimator of θ.

Theorem 5.1: Under the above assumptions of linear regression, the LSEs $\hat{\beta}_0$ and $\hat{\beta}_1$ are BLUE of β_0 and β_1, respectively.

A Few Definitions

- Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \; i = 1, \ldots, n.$
- Residuals: $e_i = y_i - \hat{y}_i, \; i = 1, \ldots, n.$
- The objective function evaluated at the LSEs is called the residual sum of squares (SS_{Res}).

$$SS_{Res} = Q(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 .$$

- The following quantity is called total sum of squares (SS_T):

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

- The following quantity is called regression sum of squares (SS_{Reg}):

$$SS_{Reg} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2 .$$

Properties of Least Squares Fit

- $\sum_{i=1}^{n} e_i = 0.$
- $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i.$
- $\sum_{i=1}^{n} x_i e_i = 0.$
- $\sum_{i=1}^{n} \hat{y}_i e_i = 0.$
- $SS_T = SS_{Reg} + SS_{Res}.$

Estimation of Error Variance

- It can be shown that

$$E(SS_{Res}) = (n-2)\sigma^2.$$

- Hence, $\hat{\sigma}^2 = \frac{SS_{Res}}{n-2} = MS_{Res}$ is an unbiased estimator of σ^2. Here MS_{Res} is the residual mean square.
- Observed value of $\hat{\sigma}^2 = \frac{SS_{Res}}{n-2}$ is called Residual variance. It's square root is called the residual standard error.
- A convenient computing formula for SS_{Res} is

$$SS_{Res} = SS_T - \hat{\beta}_1 S_{xy}.$$

Another Assumption

- Errors (ϵ_i) are normally distributed

This assumption is needed for further analysis – Hypothesis testing, construction of confidence intervals.

Hypothesis Testing: β_1

- Want to test the hypothesis that the slope parameter (β_1) equals to a constant (a value, say β_{10}):

$$H_0 : \beta_1 = \beta_{10} \text{ ag. } H_1 : \beta_1 \neq \beta_{10}$$

- Note that, $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ and y_i's are independent.
- $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}}) \implies z = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/S_{xx}}} \sim N(0, 1)$. But σ is unknown.
- $\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$. Also MS_{Res} and $\hat{\beta}_1$ are independent.
- Therefore, the test statistic is

$$t = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{Res}/S_{xx}}} \sim t_{n-2}, \text{ under } H_0.$$

- Reject H_0 iff $|t| > t_{n-2,\alpha/2}$; (at level α).

F-Test for Regression

- To test

$$H_0 : E(Y|x) = \beta_0 \text{ ag. } H_1 : E(Y|X = x) = \beta_0 + \beta_1 x$$

- Test statistics is a ratio, defined as F :

$$F = \frac{SS_{Reg}/1}{\hat{\sigma}^2} = \frac{SS_{Reg}/1}{SS_{Res}/(n-2)} \sim F_{1,n-2},$$

where $SS_{Reg} = \sum_i (\hat{y}_i - \bar{y})^2$

Hypothesis Testing: β_0

- Want to test the hypothesis that the **intercept** parameter (β_0) equals to a constant (a value, say β_{00}):

$$H_0 : \beta_0 = \beta_{00} \text{ ag. } H_1 : \beta_0 \neq \beta_{00}$$

- $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})) \implies z = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} \sim N(0, 1)$. But σ is unknown.
- $\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$. Also MS_{Res} and $\hat{\beta}_1$ are independent.
- Therefore, the test statistic is

$$t = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{Res}(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} \sim t_{n-2}, \text{ under } H_0.$$

- Reject H_0 iff $|t| > t_{n-2,\alpha/2}$; (at level α).

Interval Estimation: β_0 and β_1

- To get the CI for β_0 and β_1, the pivots are

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{MS_{Res}(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}}, \text{ and } \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MS_{Res}/S_{xx}}}, \text{ respectively.}$$

- A $100(1-\alpha)\%$ CI for β_0 is

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

- A $100(1-\alpha)\%$ CI for β_1 is

$$\left[\hat{\beta}_1 \mp t_{n-2,\alpha/2}\sqrt{\frac{MS_{Res}}{S_{xx}}}\right].$$

Interval Estimation: CI for σ^2

- To get the CI for σ^2, the pivots is

$$\frac{(n-2)MS_{Res}}{\sigma^2} \sim \chi^2_{n-2}$$

- A $100(1-\alpha)\%$ CI for σ^2 is

$$\left[\frac{(n-2)MS_{Res}}{\chi^2_{n-2;\alpha/2}}, \frac{(n-2)MS_{Res}}{\chi^2_{n-2;1-\alpha/2}}\right].$$

Interval Estimation: CI for mean response

- A regression model can be used to estimate the mean response $E(y)$ for a particular value of the regressor variable x. Let x_0 be a value of the regressor variable. Then $E(y|x_0) = \beta_0 + \beta_1 x_0$.
- Then, $\widehat{y_0} = \widehat{E(y|x_0)} = \hat{\beta}_0 + \hat{\beta}_1 x_0$
- And, $\widehat{y_0} \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)\right)$
- Pivot: $\frac{\widehat{y_0}-(\beta_0+\beta_1 x_0)}{\sqrt{MS_{Res}\left(\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$
- A $100(1-\alpha)\%$ CI for $\beta_0 + \beta_1 x_0$ is

$$\left[\widehat{y_0} \mp t_{n-2;\alpha/2}\sqrt{MS_{Res}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right].$$

Prediction Interval for New Observation:

- Let x_0 be a value of the regressor variable.
- The true value of the response is y_0 (corresponding to x_0).
- We want to provide an interval I such that $P(y_0 \in I) = 1 - \alpha$
- Note that the point estimate of y_0 is $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.
- Consider $\psi = y_0 - \hat{y}_0$.
- Then, $E(\psi) = 0$, $Var(\psi) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)$
- $\psi \sim N\left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\right)$
- Pivot: $\frac{y_0 - \hat{y}_0}{\sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$
- A $100(1-\alpha)\%$ prediction interval is

$$\left[\hat{y}_0 \mp t_{n-2;\alpha/2} \sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}\right].$$

Coefficient of determination: R^2

- Coefficient of determination is given by

$$R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{SS_{Res}}{SS_T} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2},$$

- It is a bounded quantity: $0 \leq R^2 \leq 1$.
- R^2 is interpreted as the proportion of variation explained by the model.
- Higher values of R^2 are desirable (R^2 close to 1 indicates a good fit).
- But "how high is high?": depends on the context.

Simple Linear Regression in Heights Data

- Data on heights of $n = 1375$ mothers in the UK under the age of 65 and one of their adult daughters over the age of 18 (collected and organized during the period 1893–1898 by the famous statistician Karl Pearson).
- A historical use of regression to study inheritance of height from generation to generation.
- Let's fit linear regression with this data set using `R`.