Chapter 2: Estimation

Regression Analysis

Notations

- y: response, output
- $x = (x_1, x_2, \ldots, x_p)$: predictors, input

Goal: model the relationship between y and $x_1, \ldots, x_p$

Regression Analysis Continued

- General form: $y = f(x) + \epsilon$
- $f(\cdot)$: underlying truth. Unknown
  - Linear: $f(x) = \beta_0 + \beta_1 x$
  - Polynomial: $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$
  - More complicated: $f(x) = \beta_0(cos(\beta_1 x) + sin(\beta_2 x))$
  - Non-parametric: No β's
- ϵ: error
  - We usually assume identical distributed and independent (i.i.d) errors.
  - We usually assume normally distributed errors.

Regression Analysis Continued

Underlying Condition

- We have i.i.d. n samples with p variables.

$$(x_{11}, \dots, x_{1p}, y_1), (x_{21}, \dots, x_{2p}, y_2), \cdots, (x_{n1} \dots, x_{np}, y_n)$$

Linear Regression Analysis

- There is no way to estimate $f(\cdot)$ directly given a finite number of samples.
- We put some restrictions/structure on $f(\cdot)$.
- Assume

$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

  where β_j's are unknown parameters and β_0 is the intercept.
- Benefits: Estimation of $f(\cdot)$ is reduced to estimation of β_j's

What Does "Linear" Mean?

A linear model is linear in parameters, not linear in predictors.

Formally, a function g is linear in β if

$$g(a \cdot \beta + a^* \cdot \beta^*) = a \cdot g(\beta) + a^* \cdot g(\beta^*)$$

where $a, a^* \in \mathbb{R}, \beta, \beta^* \in \mathbb{R}^p$.

Example:

- $f(x; \beta) = \beta_0 + \beta_1 e^{x_1}$ is a linear model.
- $f(x; \beta) = \beta_0 + x_1^{\beta_1}$ is not.

Linear Regression Analysis Continued

- $f(x;\beta) := \beta_0 + \beta_1 e^{x_1}$ is a linear model.
  Suppose that $\beta = (\beta_0, \beta_1)$ and $\beta^* = (\beta_0^*, \beta_1^*)$

$$
\begin{aligned}
f(x; a\beta + a^*\beta^*) &= a\beta_0 + a^*\beta_0^* + (a\beta_1 + a^*\beta_1^*)e^{x_1} \\
&= a\beta_0 + a\beta_1 e^{x_1} + \beta_0^* + a^*\beta_1^* e^{x_1} \\
&= f(x; a\beta) + f(x; a^*\beta^*)
\end{aligned}
$$

Linear Regression Analysis Continued

- $f(x;\beta) = \beta_0 + x_1^{\beta_1}$ is a not linear model.
  Suppose that $\beta = (\beta_0, \beta_1)$ and $\beta^* = (\beta_0^*, \beta_1^*)$

$$f(x; a\beta + a^*\beta^*) = a\beta_0 + a^*\beta_0^* + x_1^{a\beta_1 + a^*\beta_1^*}$$

  Since $x_1^{a\beta_1 + a^*\beta_1^*} \neq x_1^{a\beta_1} + x_1^{a^*\beta_1^*}$,

$$\begin{aligned} &\neq a\beta_0 + x_1^{a\beta_1} + \beta_0^* + x_1^{a^*\beta_1^*} \\ &= f(x; a\beta) + f(x; a^*\beta^*) \end{aligned}$$

Transformation

$f(x) = \beta_0 x_1^{\beta_1}$ is not a linear model. However, notice that

$$\ln f(x) = \ln \beta_0 + \beta_1 \ln x_1$$

Hence if we let $f^*(x) = \ln f(x), \beta_0^* = \ln \beta_0, \beta_1^* = \beta_1$, we have

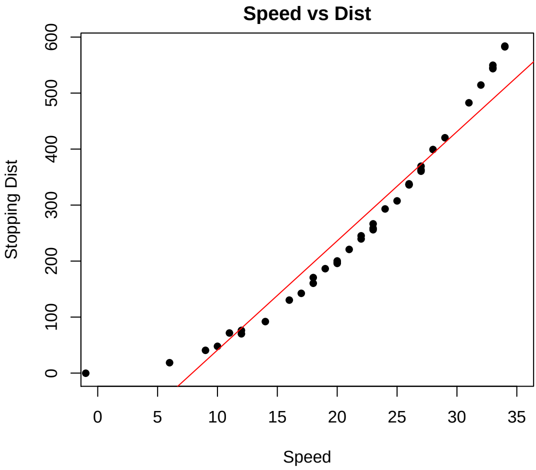
$$f^*(x) = \beta_0^* + \beta_1^* \ln x_1$$

which is a linear model.

Implications

- Linear models are less restrictive than you might think
- They can be made very flexible by transformation of the response and the predictors.

Example: Speed v.s. Stopping Distance



- Linear models are not necessarily straight lines.

Simple Linear Regression

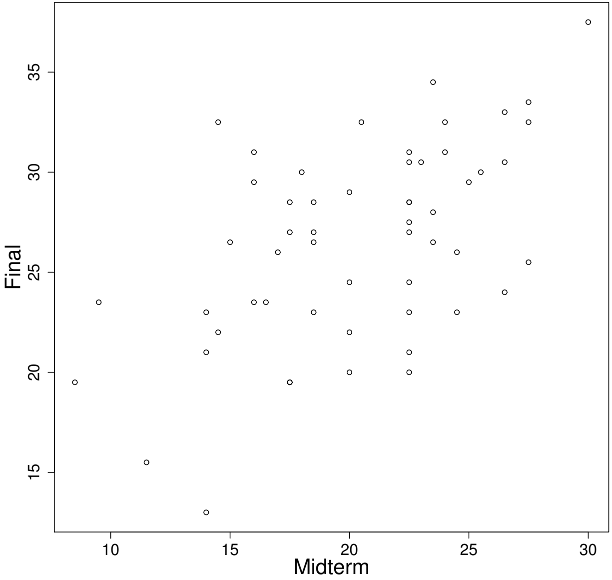
- $p = 1$, only one predictor variable
- The model is:

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

Example

- Scores from previous Stats 500
- y: final score
- x: midterm score
- $y = \beta_0 + \beta_1 x + \epsilon$

Stats 500 Data: Scatter Plot



Simple Linear Regression

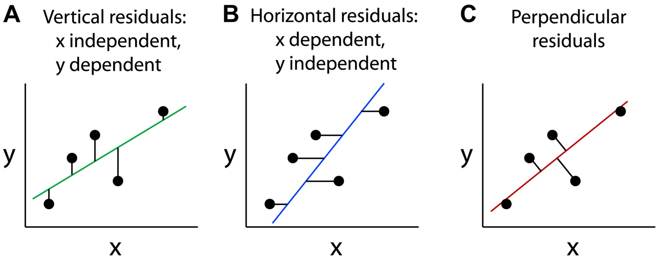
- Goal: given $(y_i, x_i),\ i = 1, \ldots, n$, estimate parameters β_0, β_1
- ϵ_i is the error term; always assume $E[\epsilon_i] = 0$.
- Minimize errors – how do we define that?

Three types of errors.

1. Vertical distance
2. Horizontal distance
3. shortest distance

Three types of errors.

1. Vertical distance
2. Horizontal distance
3. Shortest distance



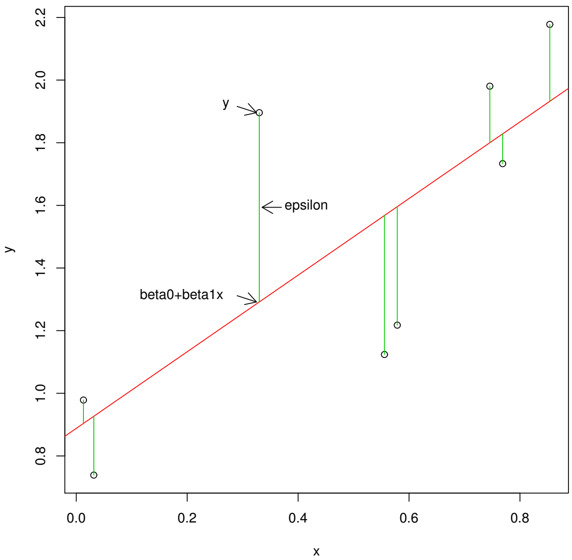
Simple Linear Regression

- One criterion is **least squares**:

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

- It minimizes vertical distances between observations and the fitted line.

Least Squares Estimate



Estimating β_0, β_1

Differentiate the criterion with respect to β_0, β_1 and set the derivatives equal to 0, we get:

$$\begin{aligned}
\frac{\partial}{\partial \beta_0} &= (-2) \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\
\frac{\partial}{\partial \beta_1} &= (-2) \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0
\end{aligned}$$

Solving for β_0 and β_1, we have:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

"Hat" notation is used for estimates.

Another interpretation

Letting $r = Cor(x, y)$, $s_y = SD(y)$, $s_x = SD(x)$, can rewrite the line equation (simple algebra) as

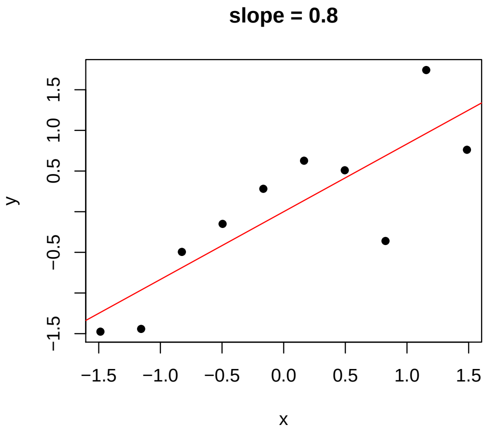
$$\frac{y - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x},$$

or, if x and y are standardized first (mean 0, sd 1), simply

$$y = rx.$$

Question

Suppose that x and y have both been standardized. In addition $cor(x, y) = 0.8$. Then, what if we regress x on y?: $x = \beta_0 + \beta_1 y$



Multiple Linear Regression

Model: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$

- # of predictors = p
- # of parameters = p + 1

Assume $E(\epsilon_i) = 0, \quad i = 1, \ldots, n$

Matrix Notation

Let

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & x_{ij} & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Then we can write the model for the data as:

$$y_{n\times 1} = X_{n\times(p+1)}\beta_{(p+1)\times 1} + \epsilon_{n\times 1}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & x_{ij} & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

This is the same model in more compact notation.

Estimating β

- Observe y and X (n i.i.d. samples)
- Want to minimize errors (vertical errors)
- Least squares criterion:

$$\begin{aligned}\min_{\beta}\sum_{i=1}^{n}\epsilon_i^2 &= \epsilon^T\epsilon \\ &= (y - X\beta)^T(y - X\beta) \\ &= y^Ty - 2y^TX\beta + \beta^TX^TX\beta\end{aligned}$$

Estimating β Ctd

Differentiating the criterion with respect to β and setting the derivative equal to 0:

- The **normal equation**:

$$X^T X \hat{\beta} = X^T y$$

- Solve for β:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T y$$

- X full rank $\Leftrightarrow X^T X$ invertible

Fitted Model

- Fitted values: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$
- Fitted model: $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$
- Residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$
- Residual sum of squares (RSS): $\sum_{i=1}^{n} \hat{\epsilon}_i^2$

Hat Matrix (Projection Matrix)

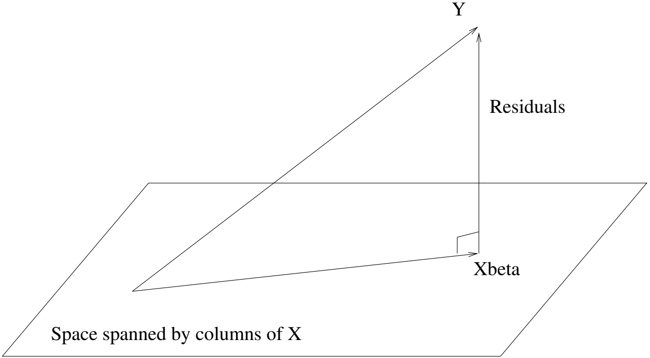
- $\hat{y} = X\hat{\beta} = X\left(X^T X\right)^{-1} X^T y = Hy$, where

$$H = X\left(X^T X\right)^{-1} X^T$$

  is called the "Hat" matrix.

- Fitted values: $\hat{y} = Hy$
- Residuals: $\hat{\epsilon} = y - \hat{y} = (I - H)y$
- The projection matrix H projects $y_{n\times 1}$ onto the column space of $X_{n\times(p+1)}$, which leads to the vector space interpretation of least squares estimate.

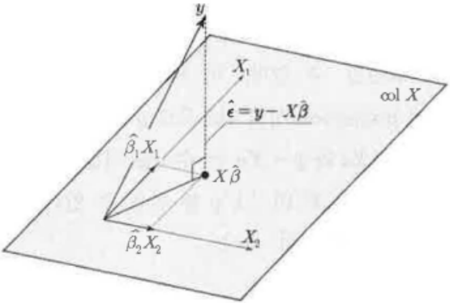
Vector Space Interpretation



$\min_{\beta}(y - X\beta)^T(y - X\beta)$ minimizes the Euclidean distance between y and the linear space spanned by the columns of X.

정사영 (Orthogonal Projection)

- 벡터공간 **S** 에 포함되지 않는 임의의 벡터 $y \in \mathbb{R}^n$ 가 있을 때, **S** 에 포함되는 벡터 중 y 와 가장 가까운 벡터는 정사영 (Orthogonal Projection)으로 구할 수 있다.



정사영 (Orthogonal Projection)

- 선형독립인 m 개의 벡터 $x_1, ..., x_m \in \mathbb{R}^n$ $(m < n)$ 으로 이루어진 행렬 $X = (x_1, x_2, ..., x_m)$이 있고, X의 열벡터 공간을 $\mathbf{S} \in \mathbb{R}^n$ 라고 하자.
- 공간 **S**위에서 y 와 가장 가까운 벡터역시 기저의 선형결합으로 표현가능:

$$a_1x_1 + a_2x_2 + ... + a_mx_m = Xa.$$

- 이 벡터 Xa 는 y를 벡터공간 **S**로 정사영시켜 얻는다고 한다.

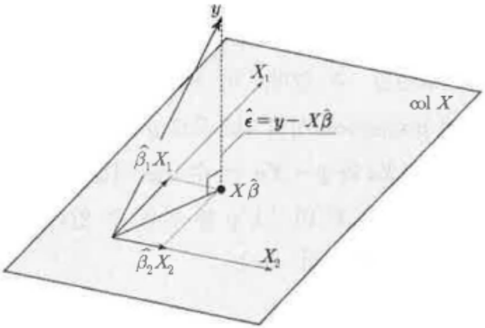
정사영 (Orthogonal Projection)

- 아래 그림처럼 $y = Xa + (y - Xa)$ 이고, Xa와 $(y - Xa)$는 수직:

$$(Xa)^T(y - Xa) = 0 \quad \Rightarrow \quad a = (X^TX)^{-1}X^Ty.$$

- **S** 에 포함되는 벡터 중 y 에 가장 가까운 벡터 Xa는

$$Xa = X(X^TX)^{-1}X^Ty = Hy.$$



사영행렬 (Projection Matrix)

- 사영행렬 (projection matrix): X가 생성하는 공간에 포함되는 벡터 중 y 에 가장 가까운 벡터를 만들어주는 $H = X(X^TX)^{-1}X^T$.
- 같은 원리로 $I - H$ 역시 사영행렬.
- 종합하면,

$$y = Hy + (I - H)y.$$

- 일반적으로 행렬 P 가 벡터공간 V의 부분공간 $\mathbf{S}$로 사영시키는 정사영행렬이 라면 임의의 벡터 $y \in V$ 에 대하여 Py는 부분공간 S 에 있는 벡터 중 y 에 가장 가까운 벡터.

$$\|y - Py\| \leq \|y - s\|, \qquad \forall s \in \mathbf{S}$$

Question

Suppose that

$$Y = \beta_1 X_1, \qquad X_1 = X_2$$

Then, what if we regress Y on (X_1, X_2)?

$$Y = \alpha_1 X_1 + \alpha_2 X_2.$$

Properties of $\hat{\beta}$

- Unbiased: $E(\hat{\beta}) = \beta$.

$$\begin{aligned} E(\hat{\beta}) &= E\left(\left(X^T X\right)^{-1} X^T y\right) \\ &= \left(X^T X\right)^{-1} X^T E(y) \\ &= \left(X^T X\right)^{-1} X^T (X\beta) \\ &= \left(X^T X\right)^{-1} \left(X^T X\right) \beta = \beta. \end{aligned}$$

- $\mathrm{Var}(\hat{\beta}) =?$ Assume $\mathrm{Var}(\epsilon) = \sigma^2 I$, then

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

- Therefore, $\widehat{\beta} \to \beta$.

Estimating Variance

- σ^2 can also be estimated:

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n - (p+1)},$$

where $n - (p+1)$ is the degrees of freedom.

- How do we determine the degrees of freedom?
  - Unbiased: $E(\hat{\sigma}^2) = \sigma^2$.

Galapagos Example

- Interested in how the number of species of tortoise on a Galapagos Island relates to other features of the island
- y: number of species of tortoise
- $x_1, \ldots, x_5$: area of the island, highest elevation of the island, distance from the nearest island, distance from Santa Cruz Island, area of the adjacent island

Galapagos Example

```
## Load the data
> library(faraway)
> data(gala)
## Check out the data
> gala
          Species Endemics    Area Elevation Nearest ...
Baltra         58       23   25.09       346     0.6 ...
Bartolome      31       21    1.24       109     0.6 ...
Caldwell        3        3    0.21       114     2.8 ...
Champion       25        9    0.10        46     1.9 ...
Coamano         2        1    0.05        77     1.9 ...
...
```

Galapagos Example

```
## Get the X matrix
> dim(gala)
[1] 30  7
> n = dim(gala)[1]
> p = dim(gala)[2] - 2
> x = cbind(1, as.matrix(gala[, 3:7]))
> ## Compute the inverse of (X^T X)
> xtx = t(x) %*% x
> xtxi = solve(xtx)
> beta = xtxi %*% t(x) %*% gala[,1]
```

```
> beta
                 [,1]
            7.068220709
Area       -0.023938338
Elevation   0.319464761
Nearest     0.009143961
Scruz      -0.240524230
Adjacent   -0.074804832
> ## Residual sum of squares
> rss = sum((gala[,1] - x %*% beta)^2)
> sigma2 = rss / (n - (p+1))
> sigma = sqrt(sigma2)
> sigma
[1] 60.97519
```

```
> ## Use the lm() function
> temp = lm(Species ~ Area + Elevation + Nearest
            + Scruz + Adjacent, data=gala)
> summary(temp)
```

```
Call:
lm(formula = Species ~ Area + Elevation + Nearest +
Scruz +  Adjacent,  data = gala)
Residuals:
     Min        1Q   Median        3Q       Max
-111.679  -34.898   -7.862   33.460  182.584
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.98 on 24 degrees of freedom
```

Goodness of Fit

- Need a measure of how well the model fits with the data
- Residual sum of squares (RSS): $\sum_i (y_i - \hat{y}_i)^2$
  Seems reasonable, but what about units?

Coefficient of determination (R^2)

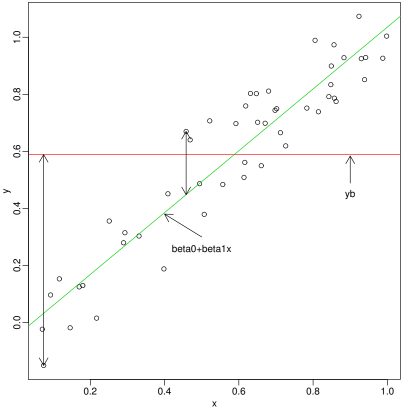
$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

- Alternative expression:

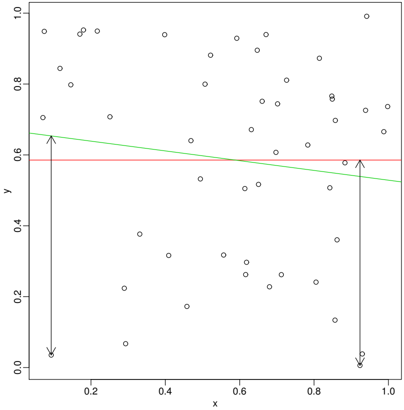
$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

- $0 \leq R^2 \leq 1$.
- R^2 “close” to 1 indicates good fit.

Intuition



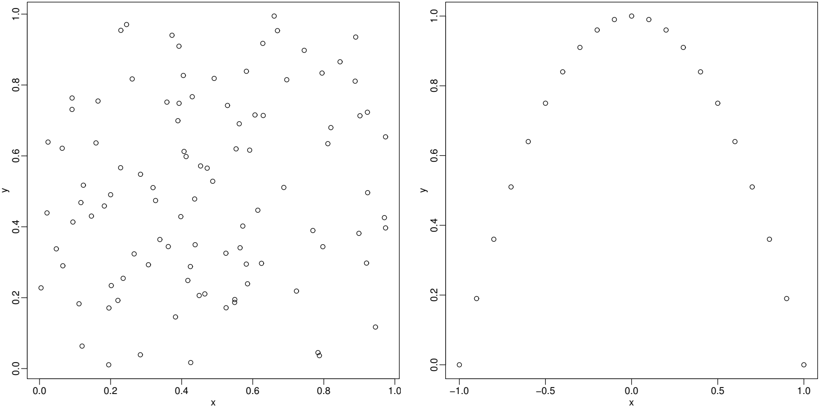
$R^2 = 0.89$



$R^2 = 0$

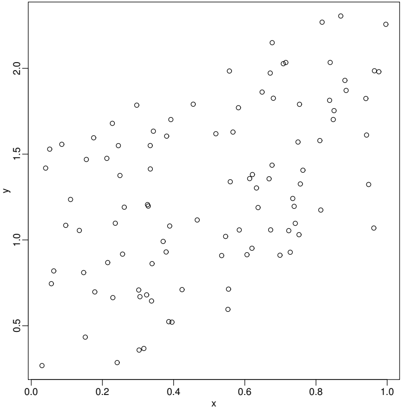
Remarks on R^2

- R^2 near 0 could be



Remarks on R^2 Continued

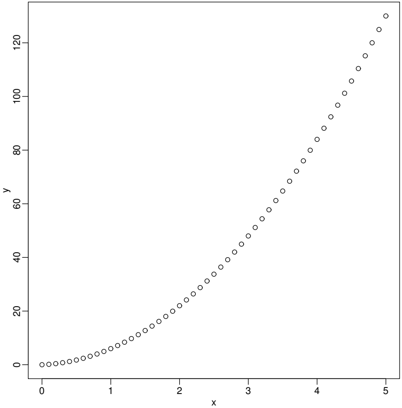
- Small R^2 does not mean that y and X are not linearly related (can have slight trend with high variance).



$R^2 = 0.2$.

Remarks on R^2 continued

- R^2 close to 1 does not mean the linear model is correct.



$$R^2 = 0.9.$$

Summary of R^2

- It may represent how well the model fits with the data.
- It may not represent how well the model fits with the data.
  - Cannot detect a linear relationship if errors are big.
  - Cannot detect a non-linear relationship
  - Prefer complicated model. (i.e., overfitting issue)

Don't trust R^2 too much

The Gauss-Markov Theorem

- Why use the least squares estimate $\hat{\beta}$?
- Theorem: Suppose $y = X\beta + \epsilon$, X is full rank, $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2 I$. Consider $\psi = c^T\beta$. Then among all unbiased linear estimates of ψ, $\hat{\psi} = c^T\hat{\beta}$ has the minimum variance and is unique.
- Example: Let $c^T = (1, x_1, \ldots, x_p)$, then $\psi = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.
- Best Linear Unbiased Estimate (BLUE)

The Gauss-Markov Theorem: Proof

Settings: Let $\tilde{\beta} = Cy$ be another linear estimator of β with

$$C = (X'X)^{-1}X' + D$$

where D is a $K \times n$ non-zero matrix. As we're restricting to unbiased estimators, minimum mean squared error implies minimum variance. The goal is therefore to show that such an estimator has a variance no smaller than that of $\widehat{\beta}$ the OLS estimator.

The Gauss-Markov Theorem: Proof

Unbiasness:

$$
\begin{aligned}
\mathsf{E}\left[\tilde{\beta}\right] &= \mathsf{E}[Cy] \\
&= \mathsf{E}\left[\left((X'X)^{-1}X' + D\right)(X\beta + \varepsilon)\right] \\
&= \left((X'X)^{-1}X' + D\right)X\beta + \left((X'X)^{-1}X' + D\right)\mathsf{E}[\varepsilon] \\
&= \left((X'X)^{-1}X' + D\right)X\beta && \mathsf{E}[\varepsilon] = 0 \\
&= (X'X)^{-1}X'X\beta + DX\beta \\
&= (I_K + DX)\beta.
\end{aligned}
$$

Therefore, $\tilde{\beta}$ is unbiased if and only if $DX = 0$. Then:

The Gauss-Markov Theorem: Proof

Variance:

$$
\begin{aligned}
\text{Var}\left(\tilde{\beta}\right) &= \text{Var}(Cy) \\
&= C \ \text{Var}(y) C' \\
&= \sigma^2 CC' \\
&= \sigma^2 \left((X'X)^{-1}X' + D\right)\left(X(X'X)^{-1} + D'\right) \\
&= \sigma^2 \left((X'X)^{-1}X'X(X'X)^{-1} + (X'X)^{-1}X'D' + DX(X'X)^{-1} + DD'\right) \\
&= \sigma^2 (X'X)^{-1} + \sigma^2 (X'X)^{-1}(DX)' + \sigma^2 DX(X'X)^{-1} + \sigma^2 DD' \\
&= \sigma^2 (X'X)^{-1} + \sigma^2 DD' \\
&= \text{Var}\left(\widehat{\beta}\right) + \sigma^2 DD'
\end{aligned}
$$

Since DD^T is a positive semi-definite matrix, $\text{Var}\left(\tilde{\beta}\right)$ exceeds $\text{Var}\left(\widehat{\beta}\right)$ by a positive semidefinite matrix.

Orthogonality

Suppose we can partition $X = [X_1 \mid X_2]$ such that $X_1^T X_2 = 0$.

Then,

$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

and

$$X^T X = \begin{pmatrix} X_1^T X_1 & 0 \\ 0 & X_2^T X_2 \end{pmatrix}.$$

Hence,

$$\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y, \qquad \hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T Y.$$

$\hat{\beta}_1$ is the same regardless of the value of X_2.

Identifiability

The least squares estimate is the solution to the normal equations:

$$(X^T X)\widehat{\beta} = X^T Y.$$

If $(X^T X)$ is singular and cannot be inverted, then there will be infinitely many solutions to the normal equations and is at least partially unidentifiable.

What Can Go Wrong?

- $X^T X$ could be singular (happens if predictors are linearly dependent or if $p > n$)
- Assumed $\mathrm{Var}(\epsilon) = \sigma^2 I$
  - Independent errors.
  - Constant and equal variance.
- Best only among linear, unbiased estimates