Predictive Analytics
Regression and Classification
Lecture 2 : Part 1

Sourish Das

Chennai Mathematical Institute

Aug-Nov, 2020

Sampling distribution of $\boldsymbol{\beta}$

- Consider the standard linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$ and $n > p$

- This implies $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$

- The least square estimator of β is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

- The sampling distribution of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1})$$

Sampling distribution of β

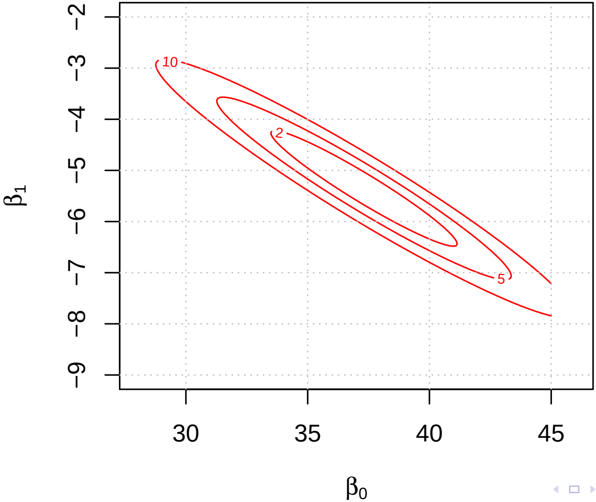
Result If $\mathbf{y}_p \sim \mathcal{N}_p(\mu, \Sigma)$, and $c_{q\times p}$ matrix. Then $\mathbf{z} = c\mathbf{y} \sim \mathcal{N}_q(c\mu, c\Sigma c^T)$

You can use this result to argue that the sampling distribution of $\hat{\beta}$ is

$$\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

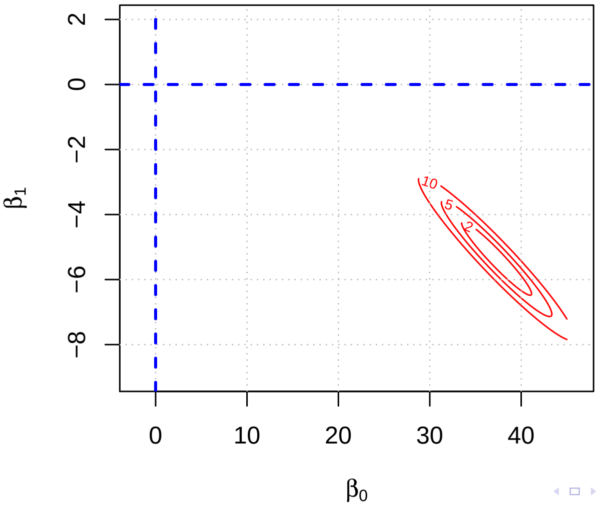
Sampling distribution for β_0 and β_1

mpg=β_0+$\beta_1$wt+ϵ



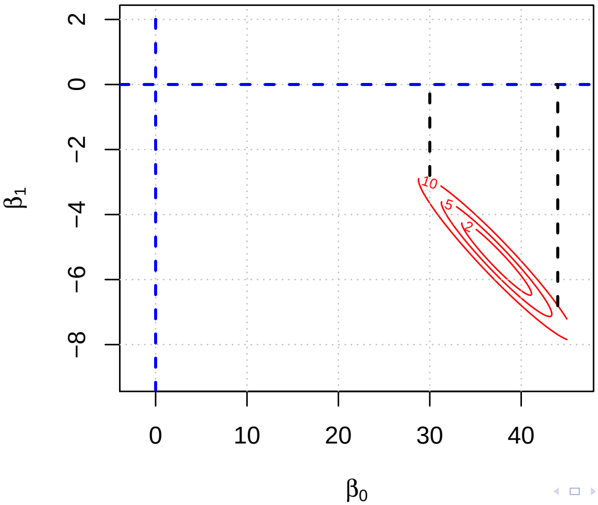
Sampling distribution for β_0 and β_1

$\texttt{mpg}=\beta_0+\beta_1\texttt{wt}+\epsilon$



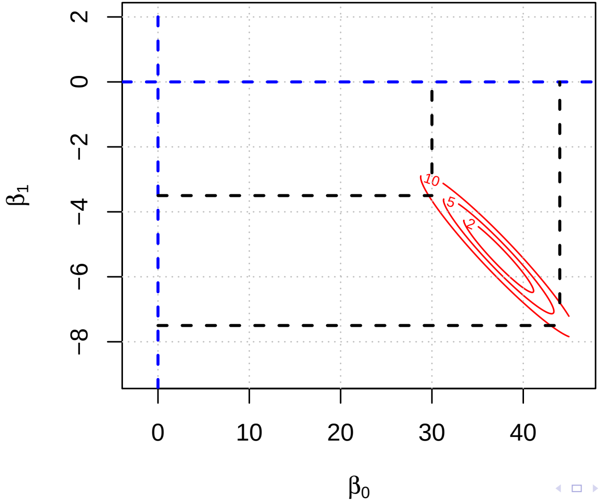
Sampling distribution for β_0 and β_1

$\texttt{mpg}=\beta_0+\beta_1\texttt{wt}+\epsilon$



Sampling distribution for β_0 and β_1

$$\texttt{mpg}=\beta_0+\beta_1\texttt{wt}+\epsilon$$



Sampling distribution

- $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$
- OLS estimator is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- Sampling distribution of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

- Residual Sum of Square is

$$RSS = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

In addition,

$$RSS \sim \sigma^2 \chi^2_{n-p}$$

Statistical Inference for β

- For i^{th} predictor,

$$\frac{\hat{\beta}_i - \beta_i}{\sigma\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim N(0,1)$$

- From the χ^2 distribution of RSS we have

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p},$$

where $s^2 = \frac{RSS}{n-p}$, this implies

$$\mathbb{E}\left(\frac{RSS}{n-p}\right) = \sigma^2,$$

i.e., s^2 is an unbiased estimator of σ^2.

Statistical Inference for $\boldsymbol{\beta}$

- Note that in the sampling distribution of $\hat{\boldsymbol{\beta}}$, the σ^2 is unknown

- As we estimate the σ^2 by its corresponding unbiased estimator $s^2 = \frac{RSS}{n-p}$,

$$t = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim t_{n-p},$$

  where $s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}$ is the standard error of $\hat{\beta}_i$

- To test null hypothesis $H_0 : \beta_i = 0$ (predictor X_i has no impact on the dependent variable y) - we can use the statistic t.

Statistical Inference for β

- To test null hypothesis $H_0 : \beta_i = 0$
  (predictor X_i has no impact on the dependent variable y)

- Alternate hypothesis $H_A : \beta_i \neq 0$
  (predictor X_i has impact on the y)

- Under the H_0, test statistics is

$$t = \frac{\hat{\beta}_i - 0}{s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} \sim t_{n-p}$$

At $100 \times \alpha\%$, level of significane, if $t_{observed} > t_{n-p}(\alpha)$ or $t_{observed} < -t_{n-p}(\alpha)$ then we reject null hypothesis.

Statistical Inference for β

- $H_0 : \beta_i = 0$ vs $H_A : \beta_i \neq 0$
- Under the H_0, test statistics is

$$t = \frac{\hat{\beta}_i - 0}{s\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{ii}}} = \frac{\hat{\beta}_i - 0}{se(\hat{\beta}_i)} \sim t_{n-p}$$

- The p-value is the probability of obtaining test results at least as extreme as the observed result, assuming that the null hypothesis is correct.

- **P-value** $= 2 * \mathbb{P}(t > |t_{oberved}||H_0$ is true$)$
- If the **P-value** is too small – we reject the null hypothesis.
- Otherwise we say we fail to reject null hypothesis

Does wt has statistically significant effect on mpg?

- mpg$=\beta_0 + \beta_1$wt$+\epsilon$

- $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.285	1.878	19.858	0
wt	-5.344	0.559	-9.559	0

- $\hat{\beta}_1 = -5.344$ and $se(\hat{\beta}_1) = 0.559$, and

$$\frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{-5.344 - 0}{0.559} = -9.559$$

and p-value < 0.01

- weight has statistically significant effect on mpg.

Does wt, and/or hp has statistically significant effect on mpg?

- mpg$=\beta_0 + \beta_1$wt$+\beta_2$hp$+\epsilon$

- $H_0 : \beta_1 = 0$ vs $H_A : \beta_1 \neq 0$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.227	1.599	23.285	0.000
wt	-3.878	0.633	-6.129	0.000
hp	-0.032	0.009	-3.519	0.001

- $\hat{\beta}_1 = -3.878$ and $se(\hat{\beta}_1) = 0.633$, and under H_0,

$$\text{t-value} = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{-3.878 - 0}{0.633} = -6.129$$

and p-value < 0.01

- weight has statistically significant effect on mpg.

Does wt, and/or hp has statistically significant effect on mpg?

- mpg$=\beta_0 + \beta_1$wt$+\beta_2$hp$+\epsilon$

- $H_0 : \beta_2 = 0$ vs $H_A : \beta_2 \neq 0$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.227	1.599	23.285	0.000
wt	-3.878	0.633	-6.129	0.000
hp	-0.032	0.009	-3.519	0.001

- $\hat{\beta}_2 = -0.032$ and $se(\hat{\beta}_2) = 0.009$, and under H_0,

$$\text{t-value} = \frac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = \frac{-0.032 - 0}{0.009} = -3.519$$

and p-value < 0.01

- hp has statistically significant effect on mpg.

Compare the two models

Model 1 $\texttt{mpg}=\beta_0 + \beta_1\texttt{wt}+\epsilon$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.285	1.878	19.858	0
wt	-5.344	0.559	-9.559	0

Model 2 $\texttt{mpg}=\beta_0 + \beta_1\texttt{wt}+\beta_2\texttt{hp}+\epsilon$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	37.227	1.599	23.285	0.000
wt	-3.878	0.633	-6.129	0.000
hp	-0.032	0.009	-3.519	0.001

1. Model 1 is a 2D model, and Model 2 is a 3D model: Are they comparable?
2. The $se(\hat{\beta}_1)$ in Model 2 is higher than Model 1. Why?

- We will discuss these issues later.

In the next part of this lecture...

- we will discuss how to check the model assumptions!
- Because if model assumptions does not hold true then any inference you do, technically those are not valid.