STA302 - Lecture 4

Cedric Beaulac

May 16, 2019

Introduction

Today's plan

- Today we introduce the linear regression model
    - The classic regression model and its assumption
    - Estimation via Maximum Likelihood
    - Inference

The linear regression model

- We will make a serie of assumptions.
- These assumptions allow us to give more information about the fitted model.
- But the results are only valid if the assumptions are respected.
- Tradeoff we make as statisticians.

The linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- Where e_i is a random variable.

The linear regression model

- Gauss-Markov assumptions (conditions) :
  - $\mathbf{E}(e_i) = 0$
  - $\text{Cor}(e_i, e_j) = 0 \ i \neq j$
  - $\text{Var}(e_i) = \sigma^2 < \inf \forall i$
- These conditions are sufficient to prove the Gauss-Markov theorem :
- The best linear unbiased estimator (BLUE) for $\beta' s$ are given by minimizing the mean square error (last week solution).
- Here, *best* means lowest variance.
- Proof : https://en.wikipedia.org/wiki/Gauss-Markov_theorem

The linear regression model

- Usually we also assume a distribution for e_i.
- The typical model formulation goes as follow.

The linear regression model

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

- With e_i are independantly ditributed $\sim N(0, \sigma^2)$
- Subtely contain all the Gauss-Markov assumption and more (stronger assumptions).

Distribution implications

- The model implies y_i is normal distributed given x_i.

$$
\begin{aligned}
\mathbf{E}(y_i|x_i) &= \mathbf{E}(\beta_0 + \beta_1 x_i + e_i) \\
&= \beta_0 + \beta_1 x_i + \mathbf{E}(e_i) \\
&= \beta_0 + \beta_1 x_i
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(y_i|x_i) &= \mathrm{Var}(\beta_0 + \beta_1 x_i + e_i) \\
&= \mathrm{Var}(e_i) \\
&= \sigma^2
\end{aligned}
$$

- Thus we have $y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.
- Or in matrix notation $\mathbf{y}|\mathbf{x} \sim N(\mathbf{X}\beta, I\sigma^2)$.
- Here we talk about $\mathbf{y}|\mathbf{x}$ because we assume $\mathbf{x}$ to be known values and $\mathbf{y}$ to be random. From now, to lighten the notation we will simple say $\mathbf{y}$.

Maximum Likelihood

- Now that we have a distribution for the response y we can estimate the parameters with another technique.
- Let's do a quick review of Maximum likelihood.

Maximum Likelihood

- Assuming x a random variable is distributed according to p_θ
- Given a data set $\{x_i | i \in (1, ..., n)\}$
- We would like to produce an estimate $\hat{\theta}$ of the parameter θ.
- $\hat{\theta}_{MLE}$ is the parameter that maximizes the probability of the observer data set :

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \; p_\theta(x_1, ..., x_n)$$

- Again, it is an optimization problem (main problems in modern stats/ML).
- Optimization and statistics go hand in hand.

Maximum Likelihood

- The likelihood is a function of θ given a data set $\mathbf{x}$.
- If x_i's are indepedent, the likelihood decomposes as a product of individual probabilities :$\mathcal{L}(\theta|\mathbf{x}) = p_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} p_\theta(x_i)$.
- Since it is complicated to derive a product it is common to look at the log-likelihood :$l(\theta|\mathbf{x}) = \log(p_\theta(x_1, ..., x_n)) = \log(\prod_{i=1}^{n} p_\theta(x_i)) = \sum_{i=1}^{n} \log(p_\theta(x_i))$.
- Maximizing likelihood and maximizing log-likelihood are equivalent (monotonically increasing function).

Maximum Likelihood

- In our regression model, we have three parameters :β_0, β_1 and σ.
- Let's estimate them with maximum likelihood.
- Since $y_i|x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ then :

$$p_\theta(y_i|x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\right)$$

Maximum Likelihood

$$
\begin{aligned}
l(\theta|x_i) &= \sum_{i=1}^{n} \log(p_\theta(y_i|x_i)) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left(-\frac{(y_i-(\beta_0+\beta_1 x_i))^2}{2\sigma^2}\right)\right) \\
&= \sum_{i=1}^{n} -\frac{1}{2}\log(2\pi) - \log(\sigma) - \frac{1}{2\sigma^2}(y_i-(\beta_0+\beta_1 x_i))^2 \\
&= -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2
\end{aligned}
$$

Maximum Likelihood

- With matrix notation :

$$l(\theta|x_i) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

- Maximizing this term with respect to β is equivalent to minimizing

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{d}{d\beta}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

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

- Exactly the same optimization from last week! Incredible!
- Both minimizing the squared error and maximizing the likelihood leads to the same estimtes $\hat{\beta}$!

Maximum Likelihood

$$l(\theta|x_i) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})$$

- Maximizing this term with respect to σ :

$$\begin{aligned} 0 &= -\frac{n}{\sigma} + \frac{1}{\sigma^3}(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ \Rightarrow \hat{\sigma}^2 &= \frac{(\mathbf{y} - \mathbf{X}\hat{\beta})^T(\mathbf{y} - \mathbf{X}\hat{\beta})}{n} \\ &= \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i))^2}{n} \end{aligned}$$

Maximum Likelihood

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i))^2}{n}$$

- This is the mean squared error!
- Reminder : It is a biased estimator of σ^2 (unbiased would be $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)$).
- For the moment σ is some kind of nuisance parameters.
- It is not interesting to us, but must be taking care of to proceed.

Inference

We love statistical models

- I love statistical models!
- We now have so much information about $\hat{\beta}$'s and $\hat{y}$.

Inference

- Inference is the action of extracting information about parameters given a data set.
- Good inference procedure considers the variability about the data: how certain are we that this parameter takes this value?

Inference

- Now that we have a model with a distibution we have more information on our estimate $\hat{\beta}$.
- Since $\mathbf{y} \sim N(\mathbf{X}\beta, I\sigma^2)$. $\mathbf{E}[\mathbf{y}] = \mathbf{X}\beta$ and $\text{Var}[\mathbf{y}] = I\sigma^2$.
- Let's use lecture 3 results on **E** and **V** of random vectors.
- Since $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, then :

$$
\begin{aligned}
\mathbf{E}[\hat{\beta}] &= \mathbf{E}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{E}[\mathbf{y}] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta = \beta
\end{aligned}
$$

- $\hat{\beta}$ is an unbiased estimator!

Inference

$$
\begin{aligned}
\mathrm{Var}[\hat{\beta}] &= \mathrm{Var}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}] \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{Var}[\mathbf{y}|\mathbf{X}]((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T I\sigma^2((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1})^T) \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}((\mathbf{X}^T\mathbf{X})^T)^{-1}) \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}
\end{aligned}
$$

Inference

IMPORTANT result from distribution theory if $\mathbf{z}$ is multivariate normal : $\mathbf{z} \sim N(\mu, \Sigma)$ and if $\mathbf{w} = \mathbf{c} + \mathbf{B}\mathbf{z}$ then :

$$\mathbf{w} \sim N(\mathbf{c} + \mathbf{B}\mu, \mathbf{B}\Sigma\mathbf{B}^T)$$

- Remember : $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
- We also know $\mathbf{y} \sim N(\mathbf{X}\beta, I\sigma^2)$.
- We can compute the distribution of $\hat{\beta}$.

Inference

$$
\begin{aligned}
&\hat{\beta} \sim N((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\beta, (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T) \\
\Rightarrow\ &\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1})\sigma^2) \\
\Rightarrow\ &\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \\
\Rightarrow\ &\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)
\end{aligned}
$$

- To begin, notice that since $\mathbf{E}(\hat{\beta}) = \beta$ we say the estimator is **unbiased**, this is an important propertie of this estimator (that's great!).
- Remember that by Gauss-Markov theorem it is BLUE (Best Linear UNBIASED estimator), i.e. amongs the unbiased linear estimator it has minimal variance. We just showed it unbiased!

Inference

- With a distribution we can also build confidence interval (small reminder : it is an interval of plausible values).
- We can then do inference; to determine if a parameters is statistically significant.
- Let's take a look at β_0 and β_1 separately.

Inference

$$
\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & . & . & 1 \\ x_1 & . & . & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ . & . \\ . & . \\ 1 & x_n \end{bmatrix}
$$

$$
= \begin{bmatrix} \sum_{i=1}^n 1^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} = n \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n}\sum_{i=1}^n x_i^2 \end{bmatrix}
$$

Inference

$$
\begin{aligned}
(\mathbf{X}^T\mathbf{X})^{-1} &= \frac{1}{\det}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \\
&= \frac{1}{n(1/n\sum_{i=1}^n (x_i^2 - (\bar{x})^2))}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}\begin{bmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}
\end{aligned}
$$

Inference

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$
$$\Rightarrow \hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{SSX}\right)$$

where $SSX = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

- $\hat{\beta}_0$ is left as an exercise.

Inference for β_1

- Since $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{SSX}\right)$, we have

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{SSX}} \sim N(0, 1)$$

- Does our predictor have a significant effect ?
- Remember $E[y_i] = \beta_0 + \beta_1 x_i$.
- If the parameter is non-zero then x affect y.
- In other word, if $\hat{\beta}_1$ is non-zero, as x moves around the expectation of y changes.
- Thus knowing the value of x helps us at better predicting y.

Inference for β_1

- Since $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{SSX}\right)$, we have

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{SSX}} \sim N(0,1)$$

- Because we don't know σ, we use $\hat{\sigma} = s$ which leads to :

$$\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{SSX}} \sim t_{n-2}$$

Inference for β_1

- The null is The paramater has no effect : $\beta_1 = 0$
- Then under the null

$$\frac{\hat{\beta}_1}{s/\sqrt{SSX}} \sim t_{n-2}$$

- If the p-value is small, this is evidence against the null.

Confidence interval

- If $z \sim N(0,1)$ then $P(-1.96 < z < 1.96) = 0.95$.
- Confidence interval are based on this concept.
- If $z \sim N(\mu, \sigma^2)$ then

$$
\begin{aligned}
P(-1.96 < \frac{z-\mu}{\sigma} < 1.96) = 0.95 \\
\Rightarrow P(-1.96\sigma < z - \mu < 1.96\sigma) = 0.95 \\
\Rightarrow P(-1.96\sigma - z < -\mu < 1.96\sigma - z) = 0.95 \\
\Rightarrow P(z - 1.96\sigma < \mu < z + 1.96\sigma) = 0.95
\end{aligned}
$$

Confidence interval

- We say a confidence interval for μ is $(z - 1.96\sigma, z + 1.96\sigma)$
- or $CI_{0.95}(\mu) = z + / - 1.96\sigma$.
- This is a confidence interval for the unknown value of μ the true parameters.
- For $x \in CI_{0.95}(\mu)$ we would accept the null $H_0 : \mu = x$ with a signficance level $1 - 0.95$.

Inference for β_1

- We have $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \frac{1}{SSX}\right)$:

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{SSX}} \sim N(0,1)$$

$$\begin{aligned}
P(-1.96 < \frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{SSX}} < 1.96) &= 0.95 \\
\Rightarrow P(-1.96\sigma/\sqrt{SSX} < \hat{\beta}_1 - \beta_1 < 1.96\sigma/\sqrt{SSX}) &= 0.95 \\
\Rightarrow P(-1.96\sigma/\sqrt{SSX} - \hat{\beta}_1 < -\beta_1 < 1.96\sigma/\sqrt{SSX} - \hat{\beta}_1) &= 0.95 \\
\Rightarrow P(\hat{\beta}_1 - 1.96\sigma/\sqrt{SSX} < \beta_1 < \hat{\beta}_1 + 1.96\sigma/\sqrt{SSX}) &= 0.95
\end{aligned}$$

Inference for β_1

- If σ is unknow. We estimate with $s = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)$
- Then we have :

$$\frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{SSX}} \sim t_{n-2}$$

Inference for β_1

- This leads to the following confidence interval for β_1 for known σ :

$$(\hat{\beta}_1 - 1.96\sigma/\sqrt{SSX}, \hat{\beta}_1 + 1.96\sigma/\sqrt{SSX})$$

and for unknown σ :

$$(\hat{\beta}_1 - t_{n-2}(\alpha/2)s/\sqrt{SSX}, \hat{\beta}_1 + t_{n-2}(\alpha/2)s/\sqrt{SSX})$$

Conclusion

- By defining a model and assuming distributions we now have a model that allows us to establish an entire interval of possible values.
- We know exactly how *sure* we are for the proposed values for $\hat{\beta}_1$.
- We can then claim if predictor x has a statistically significant effect on y.
- But what if our assumptions are violated ? Our results are less trustworthy.
- That's the next topic we will cover.

Practice problem

- Find the expectation, variance and distribution for $\hat{\beta}_0$.
- A Modern Approach to Regression with R ch.2 : 4,5,7 solutions (here)
- Alison Gibbs' additional chapter 2 practice problems (everything except 3) (here)
- A Modern Approach to Regression with R ch.2 : 1(a,b) (do with R)

External Sources

- A Modern Approach to Regression with R ch.2
- A Modern Approach to Regression with R ch.5

Test1

- Rooms : A-R BA1160, S-Y BA1170, Z BA2135 (first letter on your student ID)
- Duration : 3 Hours (ish, 2h50 minutes to be precise). I think it is a 1h30-2h00 test.
- Types of questions : Some conceptual questions, some math (expectation an such) and some R output.
- Topics :
  - p-values,statistical significance and hypothesis testing.
  - ANOVA
  - Linear Regression