Lecture 11: Linear models

STATS 101: Foundations of Statistics

Linh Tran
linh@thetahat.ai

February 12, 2019

Announcements

- Last class for this section. Will take an approx 2 week break.
- A *Colab* script is available for today's class.

Outline

Linear models

- The joint distribution
- OLS
- Regression
- Sampling distribution
- GLM
- Model selection

Recall

Given $X_n, \ldots, X_n \overset{iid}{\sim} P_0$, we can form an estimator

$$\hat{\theta}_n = \omega(X_1, \ldots, X_n) \tag{1}$$

of some underlying parameter on P_0.

- $\hat{\theta}$ has a sampling distribution
- We can try to find estimators that reach the CRLB
- We can opt for estimators that are ranges (i.e. CI's)
- We can conduct tests against a null hypothesis
- We can derive estimators that are robust against model mis-specification

Re-defining our data

Recall: for multiple random variables per observation, we have

$$P(X^1, X^2, ..., X^p) = P(X^1)(X^2|X^1) \cdots P(X^p|X^1, ..., X^{p-1}) \quad (2)$$

Re-defining our data

Recall: for multiple random variables per observation, we have

$$P(X^1, X^2, ..., X^p) = P(X^1)(X^2|X^1) \cdots P(X^p|X^1, ..., X^{p-1}) \quad (2)$$

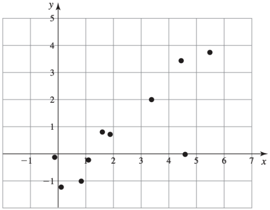
Many times, we have an *outcome* Y that we care about. Our likelihood is therefore

$$P(X^1, X^2, ..., X^p, Y) = P(X^1)(X^2|X^1) \cdots P(Y|X^1, ..., X^p) \quad (3)$$

Note: While we have the entire probability distribution to think about, we only care about $P(Y|X^1, ..., X^p)$. More specifically, we tend to look at

$$\mathbb{E}[Y|X^1, ..., X^p] \quad (4)$$

A common approach:
Let $(X_i, Y_i) : i = 1, ..., n$. Try to estimate $\mathbb{E}[Y|X]$ using a straight line.

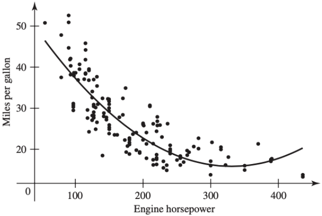


i.e. Try to find parameter values β_0, β_1 such that

$$y = \beta_0 + \beta_1 x + \epsilon \tag{5}$$

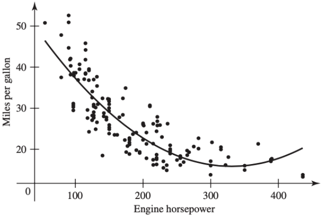
Ordinary least squares

Question: What if the relationship appears quadratic? e.g.



Ordinary least squares

Question: What if the relationship appears quadratic? e.g.



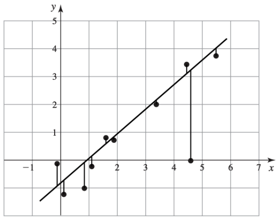
We can specify a quadratic curve (rather than linear):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon \tag{6}$$

Ordinary least squares

We can use a cost function to estimate the β_i's, i.e.

$$Q = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2 \tag{7}$$

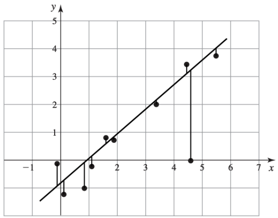


Question: How do we find the β_i's that minimize Q?

Ordinary least squares

We can use a cost function to estimate the β_i's, i.e.

$$Q = \sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i)^2 \tag{7}$$



Question: How do we find the β_i's that minimize Q?

- Calculus

Estimating β_i's

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 x_i) \tag{8}$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 x_i)\, x_i \tag{9}$$

Setting the derivatives to 0 and solving gives us:

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n \tag{10}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n \bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2} \tag{11}$$

Estimating β_i's

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 x_i) \tag{8}$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 + \beta_1 x_i)\, x_i \tag{9}$$

Setting the derivatives to 0 and solving gives us:

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n \tag{10}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2} \tag{11}$$

Question: What if we have more β's?

Estimating β_i's

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}\left(y_i - \beta_0 + \beta_1 x_i\right) \tag{8}$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}\left(y_i - \beta_0 + \beta_1 x_i\right) x_i \tag{9}$$

Setting the derivatives to 0 and solving gives us:

$$\hat{\beta}_0 = \bar{y}_n - \hat{\beta}_1 \bar{x}_n \tag{10}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}_n\bar{y}_n}{\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2} \tag{11}$$

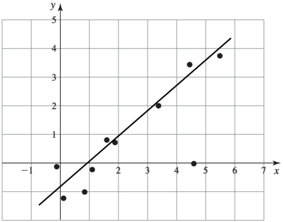
Question: What if we have more β's?

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top y \tag{12}$$

Regression

Generally, we assume a statistical model for $P(Y|X=x)$, e.g. a normal distribution

$$P(Y|X=x) = \frac{1}{2\pi\sigma^2}^{n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right) \quad (13)$$



Equivalently: we're assuming that

$$\mathbb{E}[Y|X^1 = x^1, ..., X^p = x^p] = \beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p \quad (14)$$

and var $\left(Y|X^1 = x^1, ..., X^p = x^p\right) = \sigma^2$.

Estimating β_i's

Question: How do we find the β_i's assuming a normal distribution?

Estimating β_i's

Question: How do we find the β_i's assuming a normal distribution?

We're assuming that $y_i|x_i \overset{iid}{\sim} N(\beta_0 - \beta_1 x_i, \sigma^2)$.

$$\ell(\beta, \sigma) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \quad (15)$$

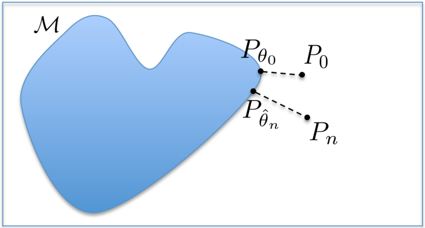
Maximum likelihood estimation:

- Take the derivative of $\ell(\beta, \sigma)$.
- Set the derivative equal to 0.
- Solve for θ.

n.b. Despite the two different approaches, the solution for β_i is equivalent between OLS and MLE.

Mis-specification

Recall: A statistical model is a set of distributions we impose on our data, i.e.



This also applies to our assumption on $P(Y|X^1, ..., X^p)$.

Sampling distribution

It can be shown that (under correct model specification) $\hat{\beta}_n$ has a normal distribution, with

$$\begin{aligned} \operatorname{var}\left(\hat{\beta}_1\right) &= \frac{\sigma^2}{s_x^2} && (16) \\ \operatorname{var}\left(\hat{\beta}_0\right) &= \sigma^2\left(\frac{1}{n}+\frac{\bar{x}_n^2}{s_x^2}\right) && (17) \\ \operatorname{cov}(\hat{\beta}_0, \hat{\beta}_1) &= -\frac{\bar{x}_n\sigma^2}{s_x^2} && (18) \end{aligned}$$

where $s_x^2 = \left(\sum_{i=1}^n (x_i - \bar{x}_n)^2\right)^{1/2}$.

Colab link

Sampling distribution

The same example in R:

```
##
## Call:
## glm(formula = y ~ x, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2484  -0.6720  -0.0138   0.7554   3.6443
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08381    0.03290   2.547    0.011 *
## x            1.00643    0.03181  31.640   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.082601)
##
##     Null deviance: 2164.2  on 999  degrees of freedom
## Residual deviance: 1080.4  on 998  degrees of freedom
## AIC: 2921.2
##
## Number of Fisher Scoring iterations: 2
```

Logistic regression

Question: What if $Y \in \{0, 1\}$?

$$\mathbb{E}[Y|X^1 = x^1, ..., X^p = x^p] = P(Y|X^1 = x^1, ..., X^p = x^p) \in [0, 1] \tag{19}$$

Allowing our model to be linear means that

$$\mathbb{E}[Y|X^1 = x^1, ..., X^p = x^p] = \beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p \in (-\infty, \infty) \tag{20}$$

Logistic regression

Question: What if $Y \in \{0, 1\}$?

$$\mathbb{E}[Y|X^1 = x^1, ..., X^p = x^p] = P(Y|X^1 = x^1, ..., X^p = x^p) \in [0, 1] \tag{19}$$

Allowing our model to be linear means that

$$\mathbb{E}[Y|X^1 = x^1, ..., X^p = x^p] = \beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p \in (-\infty, \infty) \tag{20}$$

A common solution: Use a *link function* g to restrict the outcome space, i.e.

$$\mathbb{E}[Y|X^1 = x^1, ..., X^p = x^p] = g^{-1}(\beta_0 + \beta_1 x^1 + \cdots + \beta_p x^p) \tag{21}$$

n.b.

- *Logistic regression* has $g(z) = \log\left(\frac{z}{1-z}\right)$
- logistic regression has no closed form solution

Link functions

Distribution	Support	Name	Function
Normal	$(-\infty, \infty)$	Identity	$g(z) = z$
Inverse Gaussian	$(0, \infty)$	Inverse squared	$g(z) = z^{-2}$
Poisson	$0, 1, \dots$	log	$g(z) = \log(z)$
Bernoulli	$\{0, 1\}$	logit	$g(z) = \log\left(\frac{z}{1-z}\right)$
Bernoulli	$\{0, 1\}$	probit	$g(z) = \Phi^{-1}(z)$

n.b. Estimation of the β_i's happens via MLE.

Model selection

Recall: We can use the likelihood ratio test (LRT) to do hypothesis testing.

$$\Gamma(x) = -2\log\left[\frac{\sup_{\theta\in\Theta_0} L(\theta|x_1,...,x_n)}{\sup_{\theta\in\Theta} L(\theta|x_1,...,x_n)}\right] \tag{22}$$

$$= 2\left[\ell(\hat{\theta}_{MLE}|x_1,...,x_n) - \sup_{\theta\in\Theta_0}\ell(\theta|x_1,...,x_n)\right] \tag{23}$$

Here, the hypothesis test is regarding if 1 model '*fits*' the data better than another.

Colab link

References

- DeGroot & Schervish Chapters 11.1-11.3, 11.5