STAT 331 - Applied Linear Models

Cameron Roopnarine

Last updated: September 14, 2020

Contents

LECTURE 1 | 2020-09-08

Regression model infers the relationship between:

- Response (dependent) variable: variable of primary interest, denoted by a capital letter such as Y.
- Explanatory (independent) variables: (covariates, predictors, features) variables that potentially impact response, denoted $(x_1, x_2, \ldots, x_p)$.

Alligator data:

- Y: length (m)
- x_1: male/female (categorical, 0 or 1)

Mass in stomach:

- x_2: fish
- x_3: invertebrates
- x_4: reptiles
- x_5: birds
- $x_6, \ldots, x_p$: other variables

We imagine we can explain Y in terms of $(x_1, \ldots, x_p)$ using some function so that $Y = f(x_1, \ldots, x_p)$.

In this course, we will be looking at linear models.

The Linear regression model assumes that

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

- $Y =$ value of response
- $x_1, \ldots, x_p =$ values of p explanatory variables (assumed to be fixed constants)
- $\beta_0, \beta_1, \ldots, \beta_p =$ model parameters
  - $\beta_0 =$ intercept, expected value of Y when all $x_j = 0$.
  - $\beta_1, \ldots, \beta_p$ all quantify effect on x_j on Y, $j = 1, \ldots, p$
  - $\varepsilon =$ random error

A good quote:

> "All models are wrong, but some are useful."

Assume $\varepsilon \sim N(0, \sigma^2)$. In general, the model will not perfectly explain the data.

Q: What is the distribution of Y under these assumptions?

We know:

- $\mathbf{E}[Y] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, and
- $\mathbf{Var}[Y] = \mathbf{Var}[\varepsilon] = \sigma^2$.

Therefore,

$$Y \sim N\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2\right)$$

LECTURE 2 | 2020-09-09

A linear model with response variable (Y) and *one* explanatory variable (x) is called a **simple linear regression**; that is,

$$\bar{Y} = \beta_0 + \beta_1 x + \varepsilon$$

Data consists of pairs (x_i, y_i) where $i = 1, \ldots, n$.

Before fitting any model, we might

- make a scatterplot to visualize if there is a linear relationship between x and y
- calculate *correlation*

If X and Y are random variables, then

$$\rho = \mathbf{Corr}\left[X, Y\right] = \frac{\mathbf{Cov}\left[X, Y\right]}{\mathbf{Sd}\left[X\right]\mathbf{Sd}\left[Y\right]}$$

Based on (x_i, y_i) we can estimate the sample correlation:

$$\begin{aligned} r &= \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})}} \\ &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \end{aligned}$$

The sample correlation measures the strength and direction of the *linear* relationship between X and Y.

- $|r| \approx 1$ strong linear relationship
- $|r| \approx 0$ lack of linear relationship
- $r > 0$ positive relationship
- $r < 0$ negative relationship
- $-1 \leqslant r \leqslant 1$

But does not tell us how to predict Y from X. To do so, we need to estimate β_0 and β_1.

For data (x_i, y_i) for $i = 1, \ldots, n$, the simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Assume

$$\varepsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

Therefore,

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

In other words,

$$\mathbf{E}\left[Y_i\right] = \mu_i = \beta_0 + \beta_1 x_i \text{ and } \mathbf{Var}\left[Y_i\right] = \sigma^2$$

Note that the Y_i's are independent, but they are *not* independently distributed.

Use the *Least Squares* (LS) to estimate β_0 and β_1.

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i)\right]^2 = S(\beta_0, \beta_1)$$

LS is equivalent to MLE when ε_i's are iid and Normal.

Taking partial derivatives:

$$\frac{dS}{d\beta_0} = 2\sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 x_i)\right](-1)$$

$$\frac{dS}{d\beta_1} = 2\sum_{i=1}^{n}\left[y_i - (\beta_0 + \beta_1 x_i)\right](-x_i)$$

Now,

$$\frac{dS}{d\beta_0} = 0 \iff \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0 \iff \beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\begin{aligned}
\frac{dS}{d\beta_1} = 0 &\overset{\text{plug } \beta_0}{\iff} \sum_{i=1}^{n}\left[y_i - \bar{y} + \beta_1\bar{x} - \beta_1 x_i\right] x_i = 0 \\
&\iff \sum_{i=1}^{n} x_i(y_i - \bar{y}) - \beta_1 \sum_{i=1}^{n} x_i(x_i - \bar{x}) = 0 \\
&\iff \beta_1 = \frac{\sum\limits_{i=1}^{n} x_i(y_i - \bar{y})}{\sum\limits_{i=1}^{n} x_i(x_i - \bar{x})}
\end{aligned}$$

We can also show that

$$\beta_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{x})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$$

We use a hat on the β's to show that they are estimates; that is,

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

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}$$

Call $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ the **fitted values** and $e_i = y_i - \hat{\mu}_i$ the **residual**.

Lecture 3 | 2020-09-14

Model: $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Equation of fitted line: $y = \hat{\beta}_0 + \hat{\beta}_1 x$

Interpretation:

- $\hat{\beta}_0$ is the estimate of the expected response when $x = 0$ (but not always meaningful if outside range of x_i's in data)
- $\hat{\beta}_1$ is the estimate of expected change in response for unit increase in x

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x}) \sum\limits_{i=1}^{n}(y_i - \bar{y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

- σ^2 is the "variability around the line."

  Recall that $\sigma^2 = \mathbf{Var}[\varepsilon_i] = \mathbf{Var}[Y_i]$

Q: How to estimate σ^2?

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$$
$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Intuition: use variability in residuals to estimate σ^2.

We use

$$\hat{\sigma}^2 = \frac{\sum\limits_{i=1}^{n}(e_i - \bar{e})^2}{n-2}$$

which looks looks like sample variance of e_i's. Therefore,

$$\hat{\sigma}^2 = \frac{\sum\limits_{i=1}^{n} e_i^2}{n-2} = \frac{\mathbf{Ss}\,[\text{Res}]}{n-2}$$

Note that "Square Sum" is abbreviated as "Ss". Now,

$$\bar{e} = \bar{y} - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) = 0$$

The $n-2$ will be looked are more carefully later, but for now it suffices to say that $n-2 = \text{d.f.} =$ number of parameters estimated. It allows $\hat{\sigma}^2$ to be an unbiased estimator for the true value of σ^2; that is,

$$\mathbf{E}\left[\hat{\sigma}^2\right] = \sigma^2$$

whenever $\hat{\sigma}^2$ is viewed as a random variable.

Q: Is there a statistically significant relationship?

Fact (proved using mgf in STAT 330): Suppose $Y_i \sim N(\mu_i, \sigma_i^2)$ are all independent. Then,

$$\sum_{i=1}^{n} a_i Y_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$$

for any constant a_i.

In words,

"Linear combination of Normal is Normal."

Viewing $\hat{\beta}_1$ as a random variable:

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})Y_i - \bar{Y}\sum\limits_{i=1}^{n}(x_i - \bar{x})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})x_i - \bar{x}\sum\limits_{i=1}^{n}(x_i - \bar{x})} = \frac{\sum\limits_{i=1}^{n}(x_i - \bar{x})Y_i}{\sum\limits_{i=1}^{n}(x_i - \bar{x})x_i}$$

So,

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$$

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

$$\begin{aligned}
\mathbf{E}\left[\hat{\beta}_1\right] &= \sum_{i=1}^{n} a_i \mathbf{E}\left[Y_i\right] \\
&= \sum_{i=1}^{n} a_i(\beta_0 + \beta_1 x_i) \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} \\
&= \frac{\beta_0 \overbrace{\sum_{i=1}^{n}(x_i - \bar{x})}^{=0} + \beta_1 \sum_{i=1}^{n} x_i(x_i - \bar{x})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} \\
&= \beta_1
\end{aligned}$$

On average, $\hat{\beta}_1$ is an unbiased estimator for β_1.

Now, we calculate the variance of $\hat{\beta}_1$:

$$\begin{aligned}
\mathbf{Var}\left[\hat{\beta}_1\right] &= \sum_{i=1}^{n} a_i^2 \mathbf{Var}\left[Y_i\right] \\
&= \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left[\sum_{i=1}^{n} x_i(x_i - \bar{x})\right]^2} \\
&= \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left[\sum_{i=1}^{n} (x_i - \bar{x})^2\right]^2} \\
&= \frac{\sigma^2}{S_{xx}}
\end{aligned}$$

So, since $\hat{\beta}_1$ is a linear combination of Normals,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

In a similar manner,

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates.

Then,

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

However, σ is unknown, so need to estimate with $\hat{\sigma}$:

$$\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t(n-2)$$

Since $\mathbf{Sd}\left[\hat{\beta}_1\right] = \sigma^2/S_{xx}$, we say the standard error of $\hat{\beta}_1$ is $\mathbf{Se}\left[\hat{\beta}_1\right] = \hat{\sigma}/\sqrt{S_{xx}}$

DEFINITION 0.0.1: Student's T-distribution

T is said to follow a **Student's T-distribution** with k degrees of freedom, denoted $T \sim t(k)$, if

$$T = \frac{Z}{\sqrt{U/k}}$$

where $Z \sim N(0,1)$ and $U \sim \chi^2(k)$.

Fact: For the simple linear regression model,

$$\frac{\hat{\sigma}^2(n-2)}{\sigma^2} = \frac{\mathbf{Ss}\left[\mathrm{Res}\right]}{\sigma^2} \sim \chi^2(n-2)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\dfrac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\dfrac{\hat{\sigma}^2(n-2)}{\sigma^2}\left(\dfrac{1}{n-2}\right)}} \sim t(n-2)$$

A $(1-\alpha)$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm (c)\mathbf{Se}\left[\hat{\beta}_1\right]$$

where c is the $1 - \dfrac{\alpha}{2}$ quantile of $t(n-2)$; that is,

- $P(|T| \leqslant c) = 1 - \alpha$, or
- $P(T \leqslant c) = 1 - \dfrac{\alpha}{2}$

where $T \sim t(n-2)$.

Hypothesis test: H_0: $\beta = 0$ versus H_A: $\beta_1 \neq 0$.

If H_0 is true, then

$$\frac{\hat{\beta}_1 - \beta_1}{\mathbf{Se}\left[\hat{\beta}_1\right]} = \frac{\hat{\beta}_1}{\mathbf{Se}\left[\hat{\beta}_1\right]} \sim t(n-2)$$

so calculate

$$t = \frac{\hat{\beta}_1}{\mathbf{Se}\left[\hat{\beta}_1\right]}$$

and reject H_0 at level α if $|t| > c$ where c is $1 - \dfrac{\alpha}{2}$ quantile of $t(n-2)$.

$$p\text{-value} = P(|T| \geqslant |t|) = 2P(T \geqslant |t|)$$