STAT 331 - Applied Linear Models

Cameron Roopnarine

Last updated: September 28, 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 $(x_1, x_2, \dots, x_p)$.

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
- $x_6, \dots, x_p$: other variables

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

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

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

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 $i = 1, \dots, n$.

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

$$\rho = \mathsf{Corr}(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\mathsf{Sd}(X)\mathsf{Sd}(Y)}$$

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 $i = 1, \dots, n$, 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,

$$\mathsf{E}[Y_i] = \mu_i = \beta_0 + \beta_1 x_i \text{ and } \mathsf{Var}(Y_i) = \sigma^2$$

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.

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

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,

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

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

We can also show that

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

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

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

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

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

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

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

  Recall that $\sigma^2 = \mathsf{Var}(\varepsilon_i) = \mathsf{Var}(Y_i)$

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

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

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

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\mathsf{Ss(Res)}}{n-2}$$

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 $n-2 =$ d.f. $= n -$ number of parameters estimated. It allows $\hat{\sigma}^2$ to be an unbiased estimator for the true value of σ^2; that is,

$$\mathsf{E}[\hat{\sigma}^2] = \sigma^2$$

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:

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

So,

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

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

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

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

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

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

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 $\mathsf{Sd}(\hat{\beta}_1) = \hat{\sigma}^2/S_{xx}$, we say the standard error of $\hat{\beta}_1$ is $\mathsf{Se}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{S_{xx}}$

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,

$$\frac{\hat{\sigma}^2(n-2)}{\sigma^2} = \frac{\mathsf{Ss(Res)}}{\sigma^2} \sim \chi^2(n-2)$$

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

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

$$\hat{\beta}_1 \pm (c)\mathsf{Se}(\hat{\beta}_1)$$

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

$$\frac{\hat{\beta}_1 - \beta_1}{\mathsf{Se}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\mathsf{Se}(\hat{\beta}_1)} \sim t(n-2)$$

so calculate

$$t = \frac{\hat{\beta}_1}{\mathsf{Se}(\hat{\beta}_1)}$$

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

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

Prediction for SLR: Suppose we want to predict the response y for a new value of x. Say $x = x_0$. Then, SLR model says

$$Y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$$

where Y_0 is a r.v. for response when $x = x_0$.

The fitted model predicts the *value* of y to be

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

As a random variable,

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

then,

$$\mathsf{E}[\hat{Y}_0] = \mathsf{E}[\hat{\beta}_0] + x_0 \mathsf{E}[\hat{\beta}_1] = \beta_0 + \beta_1 x_0 = \mathsf{E}[Y_0]$$

since $\hat{\beta}_i$ for $i = 0, 1$ are unbiased. We can say that $\hat{Y}_0$ is an unbiased estimate of the random variable for the mean of Y_0.

We claim that:

$$\mathsf{Var}(\hat{Y}_0) = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

by expressing $\hat{Y}_0 = \sum_{i=1}^{n} a_i Y_i$. This implies that,

$$\hat{Y}_0 \sim N \left(\beta_0 + \beta_1 x_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \right)$$

The random variable for prediction error is

$$Y_0 - \hat{Y}_0$$

where Y_0 and $\hat{Y}_0$ are independent and $\hat{Y}_0$ is a function of $Y_1, \ldots, Y_n$.

$$\mathsf{E}[Y_0 - \hat{Y}_0] = \mathsf{E}[Y_0] - \mathsf{E}[\hat{Y}_0] = 0$$

$$\mathsf{Var}(Y_0 - \hat{Y}_0) = \mathsf{Var}(Y_0) + (-1)^2 \mathsf{Var}(\hat{Y}_0) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Again, we have a linear combination of independent Normals, so

$$Y_0 - \hat{Y}_0 \sim N \left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \right)$$

Since σ is unknown, we use $\hat{\sigma}$ and get the following:

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

Intuition for prediction error composed of 2 terms:

- $\mathsf{Var}(Y_0)$: random error of new observation
- $\mathsf{Var}(\hat{Y}_0)$ (predictor): estimating β_0 and β_1

Those are 2 sources of uncertainty.

Note: Be careful that the prediction may not make sense if x_0 is outside the range of the x_i's in the data.

$(1-\alpha)$ prediction interval for y_0:

$$\hat{y}_0 \pm c\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}}$$

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

Orange production 2018 in FL

- x: acres
- y: # boxes of oranges (thousands)
- (x_i, y_i) recorded for each of 25 FL counties
- $r = 0.964$
- $\bar{x} = 16133$
- $\bar{y} = 1798$
- $S_{xx} = 1.245 \times 10^{10}$
- $S_{xy} = 1.453 \times 10^{9}$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.1167$$

which is a positive slope (positive correlation between x and y). The expected number of boxes produced is estimated to be about 117 higher per an additional acre.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -85.3$$

Not meaningful to interpret, since it is the expected production if there were 0 acres (outside the range of x_i) as no county has $x = 0$.

Now suppose

$$\mathsf{Ss}(\text{Res}) = 1.31 \times 10^7$$

the residuals are the differences between y_i and the fitted regression line.

- $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2} = \dfrac{1.31 \times 10^7}{25-2} = 5.7 \times 10^5$
- $\mathsf{Se}(\hat{\beta}_1) = \dfrac{\hat{\sigma}}{\sqrt{S_{xx}}} = 0.00676$
- To test H_0: $\beta_1 = 0$, calculate

  $$t = \frac{\hat{\beta}_1 - 0}{\mathsf{Se}(\hat{\beta}_1)} = \frac{0.1167}{0.00676} \approx 17.3$$

  Select the 0.975 quantile (for demonstration purposes) of $t(23)$ is 2.07.
- Note that 17.3 is very unlikely to see in $t(23)$.

Since $17.3 > 2.07$, we reject H_0 at $\alpha = 0.05$ level, conclude there's a significant linear relationship between acres and oranges produced.

The 95% confidence interval for β_1 is

$$0.1167 \pm 2.07(0.00676)$$

which does not contain 0.

$$p\text{-value} = P(|t_{23}| \geqslant 17.3) = 2P(t_{23} \geqslant 17.3) \approx 1.2 \times 10^{-14}$$

Predict the # of boxes in thousands produced if we had 10000 acres to grow oranges.

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = -85.3 + (0.1167)(10000) \approx 1082$$

The 95% prediction interval is:

$$1082 \pm 2.07\sqrt{5.69 \times 10^5}\sqrt{1 + \frac{1}{25} + \frac{(6133)^2}{1.245 \times 10^{10}}}$$

Note: **not** trying to establish causation.

Check LEARN for `florange.csv`.

Q: Is σ the same for all values of y?

A: It appears to not in the sense that the variance appears to be higher with respect to higher acres. Sigma will be smaller when there's less acres. Later, this will be testing equal variance or homoscedastic assumption. Later, when we talk about variable transformations we can consider taking the log.

Q: Are the error terms plausibly independent? In other words, does knowing one e_i (residual) help predict e_j (another residual) for a different county?

A: There's diagnostics for checking this. However, intuitively there could be some common factors at play when two counties are geographically close.

LECTURE 5 | 2020-09-21

Multiple Linear Regression (MLR)

p explanatory variables which can be categorical, continuous, etc.

Rocket

- x_1: nozzle area (large or small, 0 or 1)
- x_2: mixture in propellent, ratio oxidized fuel
- Y: thrust

Want to develop linear relationship between response y and $x_1, x_2, \dots, x_p$.

Data n observations, each consists of response and p explanatory variables $(y_i, x_{i1}, x_{i2}, \dots, x_{ip})$. Then,

$$Y_i \sim N(\underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}}_{\mathsf{E}[Y_i] = \mu_i}, \sigma^2)$$

or $Y_i = \mu_i + \varepsilon_i$ where $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

We can write in vector/matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \cdots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \cdots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \cdots + \beta_p x_{np} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$$

Which we can write as

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} & x_{2p} \\ \vdots & & & \ddots & & \vdots \\ 1 & x_{(n-1)1} & x_{(n-1)2} & \cdots & x_{(n-1)(p-1)} & x_{(n-1)p} \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} & x_{np} \end{bmatrix}_{n\times(p+1)}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1)\times 1}$$

We call $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)^\top$ a **random vector** (vector of r.v.'s), analogue of expectation and variance properties.

- Mean vector:

$$\mathsf{E}[\boldsymbol{Y}] = \begin{bmatrix} \mathsf{E}[Y_1] \\ \mathsf{E}[Y_2] \\ \vdots \\ \mathsf{E}[Y_n] \end{bmatrix}$$

- Covariance matrix (variance-covariance matrix):

$$\mathsf{Var}(\boldsymbol{Y}) = \begin{bmatrix} \mathsf{Var}(Y_1) & \mathsf{Cov}(Y_1, Y_2) & \cdots & \mathsf{Cov}(Y_1, Y_{n-1}) & \mathsf{Cov}(Y_1, Y_n) \\ \mathsf{Cov}(Y_2, Y_1) & \mathsf{Var}(Y_2) & \cdots & \mathsf{Cov}(Y_2, Y_{n-1}) & \mathsf{Cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathsf{Cov}(Y_{n-1}, Y_1) & \mathsf{Cov}(Y_{n-1}, Y_2) & \cdots & \mathsf{Var}(Y_{n-1}) & \mathsf{Cov}(Y_{n-1}, Y_n) \\ \mathsf{Cov}(Y_n, Y_1) & \mathsf{Cov}(Y_n, Y_2) & \cdots & \mathsf{Cov}(Y_n, Y_{n-1}) & \mathsf{Var}(Y_n) \end{bmatrix}$$

  - symmetric since $\mathsf{Cov}(Y_i, Y_j) = \mathsf{Cov}(Y_j, Y_i)$; that is $\mathsf{Var}(\boldsymbol{Y})^\top = \mathsf{Var}(\boldsymbol{Y})$.
  - positive semi-definite since $\boldsymbol{a}^\top \mathsf{Var}(\boldsymbol{Y})\boldsymbol{a} \geqslant 0$ for all $\boldsymbol{a} \in \mathbb{R}^n$.
  - $\mathsf{Var}(\boldsymbol{Y}) = \mathsf{E}[(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}])\,(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}])^\top] = \mathsf{E}[\boldsymbol{Y}\boldsymbol{Y}^\top] - \mathsf{E}[\boldsymbol{Y}]\mathsf{E}[\boldsymbol{Y}]^\top$

Properties of random vector: let $\boldsymbol{a}$ be a $1 \times n$ matrix (row vector) of constants and A be an $n \times n$ matrix of constants.

$$\mathsf{E}[\boldsymbol{a}\boldsymbol{Y}] = \boldsymbol{a}\boldsymbol{Y}$$
$$\mathsf{E}[A\boldsymbol{Y}] = A\mathsf{E}[\boldsymbol{Y}]$$
$$\mathsf{Var}(\boldsymbol{a}\boldsymbol{Y}) = \boldsymbol{a}\mathsf{Var}(\boldsymbol{Y})\boldsymbol{a}^\top$$
$$\mathsf{Var}(A\boldsymbol{Y}) = A\mathsf{Var}(\boldsymbol{Y})A^\top$$

Derivation of (4):

$$\begin{aligned} \mathsf{Var}(A\boldsymbol{Y}) &= \mathsf{E}[(A\boldsymbol{Y} - \mathsf{E}[A\boldsymbol{Y}])\,(A\boldsymbol{Y} - \mathsf{E}[A\boldsymbol{Y}])^\top] \\ &= \mathsf{E}[(A\boldsymbol{Y} - A\mathsf{E}[\boldsymbol{Y}])\,(A\boldsymbol{Y} - A\mathsf{E}[\boldsymbol{Y}])^\top] \\ &= \mathsf{E}[A\,(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}])\,(A\,(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}]))^\top] \\ &= \mathsf{E}[A\,(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}])\,(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}])^\top\,A^\top] \\ &= A\mathsf{E}[(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}])\,(\boldsymbol{Y} - \mathsf{E}[\boldsymbol{Y}])^\top]A^\top \\ &= A\mathsf{Var}(\boldsymbol{Y})A^\top \end{aligned}$$

Numerical example: $\boldsymbol{Y} = (Y_1, Y_2, Y_3)^\top$. Suppose

$$\mathsf{E}[\boldsymbol{Y}] = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

and

$$\mathsf{Var}(Y) = \begin{bmatrix} 4 & 1/2 & -2 \\ 1/2 & 1 & 0 \\ -2 & 0 & 3 \end{bmatrix}$$

and

$$\boldsymbol{a} = \begin{bmatrix} 1 & -1 & 2 \end{bmatrix}$$

and

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Exercise:

- $\mathsf{E}[\boldsymbol{aY}]$
- $\mathsf{Var}(\boldsymbol{aY})$
- $\mathsf{E}[A\boldsymbol{Y}]$
- $\mathsf{Var}(A\boldsymbol{Y})$

Let's do the first two,

$$\mathsf{E}[\boldsymbol{aY}] = \boldsymbol{a}\mathsf{E}[\boldsymbol{Y}] = \begin{bmatrix} 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} = 1(3) - 1(1) + 2(2) = 6$$

$$\begin{aligned} \mathsf{Var}(\boldsymbol{aY}) &= \boldsymbol{a}\mathsf{Var}(\boldsymbol{Y})\boldsymbol{a}^\top \\ &= \begin{bmatrix} 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 4 & 1/2 & -2 \\ 1/2 & 1 & 0 \\ -2 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 4(1) + (1/2)(-1) - 2(2) \\ (1/2)(1) + 1(-1) + 0(2) \\ -2(1) + 0(-1) + 3(2) \end{bmatrix} \\ &= \begin{bmatrix} 1 & -1 & 2 \end{bmatrix} \begin{bmatrix} -1/2 \\ -1/2 \\ 4 \end{bmatrix} \\ &= 1(-1/2) - 1(-1/2) + 2(4) \\ &= 8 \end{aligned}$$

Multivariate normal distribution (MVN): We say that $\boldsymbol{Y} \sim \text{MVN}(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}$ = mean vector and Σ = covariance matrix. Suppose $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$.

$$f(\boldsymbol{y}; \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu})\right\}$$

where Σ^{-1} is the inverse of the covariance matrix and $|\Sigma|$ is the determinant of Σ.

Properties of MVN: Suppose $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \sim \text{MVN}(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{a}$ is a $1 \times n$ row vector of constants and A is an $n \times n$ matrix of constants.

1. Linear transformations of MVN is MVN, so

$$\boldsymbol{aY} \sim \text{MVN}(\boldsymbol{a\mu}, \boldsymbol{a}\Sigma\boldsymbol{a}^\top)$$

$$A\boldsymbol{Y} \sim \text{MVN}(A\boldsymbol{\mu}, A\Sigma A^\top)$$

2. Marginal distribution of Y_i is Normal,

$$Y_i \sim N(\mu_i, \Sigma_{ii})$$

In fact, any subset of Y_i's is MVN

3. Conditional MVN is MVN, e.g. $Y_1 \mid Y_2, \dots, Y_n$
4. Another property:
$$\mathsf{Cov}(Y_i, Y_j) = 0 \iff Y_i, Y_j \text{ independent}$$
that is, Y_i and Y_j are uncorrelated.
$$\Sigma_{ij} = 0$$

LECTURE 6 | 2020-09-23

MLR: $\boldsymbol{Y} = X\boldsymbol{B} + \boldsymbol{\varepsilon}$

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

So random vector:
$$\boldsymbol{\varepsilon} \sim \text{MVN}\left(\begin{bmatrix}0\\0\\\vdots\\0\\0\end{bmatrix}, \begin{bmatrix}\sigma^2 & 0 & \cdots & 0 & 0\\ 0 & \sigma^2 & \cdots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & \sigma^2 & 0\\ 0 & 0 & \cdots & 0 & \sigma^2\end{bmatrix} \right) = (\mathbf{0}_{n\times 1}, \sigma^2 I_{n\times n})$$

since $\mathsf{Cov}(\varepsilon_1, \varepsilon_2) = 0$ due to independence.

Thus, $\boldsymbol{Y} \sim \text{MVN}(X\boldsymbol{B}, \sigma^2 I)$.

Least squares: Define
$$S(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^{n} (y_i - \underbrace{(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}_{\mathsf{E}[Y_i] = \mu_i})^2$$

First partial:
$$\frac{\partial S}{\partial \beta_0} = \sum_{i=1}^{n} 2(y_i - \mu_u)(-1)$$

We observe that all other partials for $j = 1, \dots, p$ are:
$$\frac{\partial S}{\partial \beta_j} = \sum_{i=1}^{n} 2(y_i - \mu_i)(-x_{ij})$$

Set $\dfrac{\partial S}{\partial \beta_0} = 0$ and $\dfrac{\partial S}{\partial \beta_j} = 0$ for $j = 1, \dots, p$.

$$\begin{cases} \displaystyle\sum_{i=1}^{n} (y_i - \mu_i) \iff \mathbf{1}_{n\times n}^{\top}(\boldsymbol{y} - \boldsymbol{\mu}) = 0 \\ \displaystyle\sum_{i=1}^{n} (y_i - \mu_i) x_{ij} = 0 \iff \boldsymbol{x}_j^{\top}(\boldsymbol{y} - \boldsymbol{\mu}) = 0 \quad j = 1, \dots, p \end{cases}$$

since we recall that
$$X = \begin{bmatrix}1 & x_{11}x_{12} & \cdots & x_{1p}\\ \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np}\end{bmatrix} = \begin{bmatrix}\mathbf{1}_{n\times 1} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{p-1} & \boldsymbol{x}_p\end{bmatrix}$$

Therefore,
$$X^{\top}(\boldsymbol{y} - X\boldsymbol{B}) = 0 \iff X^{\top}\boldsymbol{y} - X^{\top}X\boldsymbol{B} = 0 \iff X^{\top}X\boldsymbol{B} = X^{\top}\boldsymbol{y} \iff \boldsymbol{B} = (X^{\top}X)^{-1}X^{\top}\boldsymbol{y}$$

assuming $X^{\top}X$ is invertible (full rank of $p+1$, or linearly independent columns).

Define residuals:

$$e_i = y_i - (\underbrace{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots \hat{\beta}_p x_{ip}}_{\text{fitted value } \mu_i})$$

or equivalently,

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{B}}, \quad \boldsymbol{e} = \boldsymbol{y} - \boldsymbol{\mu}$$

and estimate σ^2 based on e_i's

$$\sigma^2 = \frac{\mathsf{Ss(Res)}}{n-(p+1)} = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1} = \frac{\boldsymbol{e}^\top \boldsymbol{e}}{n-p-1}$$

since d.f. is $n - (\text{no. estimated parameters})$. When viewed as a random variable,

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-p-1)$$

Inference for

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^\top = (X^\top X)^{-1} X^\top \boldsymbol{Y}$$

Note that $\hat{\boldsymbol{\beta}}$ is a matrix of constants and $\boldsymbol{Y}$ is a random vector, and

$$\boldsymbol{Y} \sim \mathsf{MVN}(X\boldsymbol{\beta}, \sigma^2 I)$$

$$\begin{aligned}
\mathsf{E}[\hat{\boldsymbol{\beta}}] &= \mathsf{E}[(X^\top X)^{-1} X^\top \boldsymbol{Y}] \\
&= (X^\top X)^{-1} X^\top \mathsf{E}[\boldsymbol{Y}] \\
&= (X^\top X)^{-1} (X^\top X) \boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

That is, $\mathsf{E}[\hat{\beta}_0], \ldots, \mathsf{E}[\hat{\beta}_p] = \beta_p$ all unbiased.

$$\begin{aligned}
\mathsf{Var}((X^\top X)^{-1} X^\top) &= (X^\top X)^{-1} X^\top \mathsf{Var}(\boldsymbol{Y}) \left[(X^\top X)^{-1} X^\top\right]^\top \\
&= (X^\top X)^{-1} X^\top \sigma^2 I (X^\top)^\top \left[(X^\top X)^{-1}\right]^\top && X^\top X \text{ symmetric} \\
&= \sigma^2 (X^\top X)^{-1} (X^\top)(X^\top X)(X^\top X)^{-1}
\end{aligned}$$

$\hat{\boldsymbol{\beta}}$ is a linear transformation of $\boldsymbol{Y}$, so

$$\hat{\boldsymbol{\beta}} \sim \mathsf{MVN}(\boldsymbol{\beta}, \sigma^2 \underbrace{(X^\top X)^{-1}}_{V})$$

For a specific parameter β_j,

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 V_{jj})$$

from marginal property of MVN.

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{V_{jj}}} \sim N(0,1)$$

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\sqrt{V_{jj}}} \sim t(n-p-1)$$

We define the standard error of $\hat{\beta}_j$ as

$$\mathsf{Se}(\hat{\beta}_j) = \hat{\sigma}\sqrt{V_{jj}}$$

So, a $(1-\alpha)$ confidence interval for β_j is

$$\hat{\beta}_j \pm c\mathsf{Se}(\hat{\beta}_j)$$

where c is $(1-(\alpha/2))$ quantile of $t(n-p-1)$.

To test H_0: $\beta_j = 0$ vs H_A: $\beta_j \neq 0$, calculate t-statistic

$$t = \frac{\hat{\beta}_j}{\mathsf{Se}(\hat{\beta}_j)}$$

reject at level α if $|t| > c$ and p-value is $2P(T \geqslant |t|)$ where $T \sim t(n-p-1)$.

Interpretation of $\hat{\boldsymbol{\beta}}$: fitted linear regression model says $\widehat{\mathsf{E}[Y]}$ (estimate of the expected response) is $\hat{\beta}_0 + \cdots + \beta_1 x_1 + \hat{\beta}_p x_p$.

- $\hat{\beta}_0$ is the estimate of expected response when all explanatory variables are equal to 0.
- $\hat{\beta}_j$ is the estimated change in expected response for a unit increase in x_j, when holding all other explanatory variables constant, e.g.

$$\hat{\beta}_0 + \hat{\beta}_1(x_1+1) + \cdots + \hat{\beta}_p x_p - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p) = \hat{\beta}_1$$

Rocket example: $n = 12$

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} 473.6 \\ 16.7 \\ -1.09 \end{bmatrix} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)^\top$$

- x_1: nozzle area ($1 = L, 0 = S$)
- x_2: propellent ratio
- Y: thrust

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{12} e_i^2}{12-1-2}} = 2.655$$

Interpretation of $\hat{\boldsymbol{\beta}}$:

- $\hat{\beta}_1$ estimated change in expected thrust is 16.7 when changing small to large nozzle while holding other variables (propellent ratio) constant.
- $\hat{\beta}_2$ estimated thrust to decrease by 1.09 on average for a unit increase in propellent ratio while holding other variables (nozzle area) constant.

Given: $\mathsf{Se}(\hat{\beta}_2) = 0.94$.

Then: t-statistic for H_0: $\beta_2 = 0$ vs H_A: $\beta_2 \neq 0$ is $t = -1.09/0.94 = -1.16$

$$p\text{-value} = 2P(T \geqslant 1.16) = 0.275 \text{ from R where } T \sim t(9)$$

Do not reject H_0 (e.g. $\alpha = 0.05$), therefore propellent ratio does not significantly influence thrust.

Lecture 7 | 2020-09-28

Recall:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim \mathsf{MVN}(X\boldsymbol{\beta}, \sigma^2 I)$$

- Estimates: $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \boldsymbol{Y}$
- Fitted values: $\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}$
- Residuals: $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$

Geometric interpretation of data. Constants: $X = \begin{bmatrix}\mathbf{1} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_p\end{bmatrix}_{n\times(p+1)}$

Values of responses: $\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$

Recall: $\text{Span}(X) = \{b_0\mathbf{1} + b_1\boldsymbol{x}_1 + \cdots + b_p\boldsymbol{x}_p : b_0, \ldots, b_p \in \mathbb{R}\} \subset \mathbb{R}^n$ which is all linear combinations of columns of X which is a subspace of $\mathbb{R}^n$. Recall, by assumption $\text{rank}(X) = p + 1$.

We can say $\text{Span}(X)$ represents all possible vector values $X\boldsymbol{b}$ where $\boldsymbol{b} = (b_0, b_1, \ldots, b_p)^\top$.

Generally, $\boldsymbol{y} \notin \text{Span}(X)$, so since the linear model is an approximation, $\boldsymbol{\varepsilon}$ variability not explained by model.

Intuitively, it makes sense to choose an estimate $\hat{\boldsymbol{\beta}}$ so that $X\hat{\boldsymbol{\beta}}$ is as close to $\boldsymbol{y}$ as possible. Therefore, $\boldsymbol{e}$ must be orthogonal to $\text{Span}(X) \iff \boldsymbol{e}$ is orthogonal to all columns of X.

$$\begin{aligned} \mathbf{1}^\top \cdot (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) &= 0 \\ \boldsymbol{x}_1^\top \cdot (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) &= 0 \\ &\vdots \\ \boldsymbol{x}_p^\top \cdot (\boldsymbol{y} - \hat{\boldsymbol{\mu}}) &= 0 \end{aligned}$$

which is the same as LS estimates.

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}}, \quad \boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{\mu}}$$

Define the **hat matrix** as

$$H = X(X^\top X)^{-1}X^\top$$

Properties of H

(1) H is symmetric.

(2) H is idempotent.

(3) $I - H$ is symmetric idempotent.

$$H^\top = [X(X^\top X)^{-1}X^\top]^\top = X(X^\top X)^{-1}X^\top = H$$
$$HH = X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top = H$$

Let's view $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{e}$ as random vectors

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}} = X(X^\top X)^{-1}X^\top\boldsymbol{Y} = H\boldsymbol{Y}$$

$$\boldsymbol{e} = \boldsymbol{Y} - \hat{\boldsymbol{\mu}} = I\boldsymbol{Y} - H\boldsymbol{Y} = (I - H)\boldsymbol{Y}$$

$$\mathsf{E}[\hat{\boldsymbol{\mu}}] = \mathsf{E}[H\boldsymbol{Y}] = H\mathsf{E}[\boldsymbol{Y}] = X(X^\top X)^{-1}X^\top \underbrace{X\boldsymbol{\beta}}_{\mathsf{E}[\boldsymbol{Y}]} = X\boldsymbol{\beta}$$

$$\mathsf{Var}(\hat{\boldsymbol{\mu}}) = \mathsf{Var}(H\boldsymbol{Y}) = H\mathsf{Var}(\boldsymbol{Y})H^\top = H\sigma^2 I H^\top = \sigma^2(HH^\top) = \sigma^2 H$$

$$\mathsf{E}[\boldsymbol{e}] = \mathsf{E}[(I - H)\boldsymbol{Y}] = \mathsf{E}[\boldsymbol{Y}] - \mathsf{E}[H\boldsymbol{Y}] = X\boldsymbol{\beta} - X\boldsymbol{\beta} = 0$$

$$\mathsf{Var}(\boldsymbol{e}) = (I - H)\mathsf{Var}(I - H)^\top = \sigma^2(I - H)(I - H)^\top = \sigma^2(I - H)$$

So since $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{e}$ are linear transformations of $\boldsymbol{Y}$

$$\hat{\boldsymbol{\mu}} \sim \text{MVN}(X\boldsymbol{\beta}, \sigma^2 H)$$

$$\hat{\boldsymbol{e}} \sim \text{MVN}(0, \sigma^2(I - H))$$

Prediction: Suppose we want to predict response for (the first 1 represents the intercept)

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 & x_{01} & x_{02} & \cdots & x_{0p} \end{bmatrix}_{1\times(p+1)}$$

Let Y_0 random variable representing the response associated with $\boldsymbol{x}_0$. The MLR says

$$Y_0 \sim N(\beta_0 + \beta_1 x_{01} + \beta_p x_{0p}, \sigma^2)$$

So we predict the value

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} = \boldsymbol{x}_0 \hat{\boldsymbol{\beta}}$$

which represents the estimated mean response given $x_{01}, x_{02}, \ldots, x_{0p}$. Corresponding distribution has

$$\mathsf{E}[\hat{Y}_0] = \boldsymbol{x}_0 \mathsf{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{x}_0 \boldsymbol{\beta} = \mathsf{E}[Y_0]$$

$$\mathsf{Var}(\hat{Y}_0) = \boldsymbol{x}_0 \mathsf{Var}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}_0^\top = \boldsymbol{x}_0 \sigma^2 (X^\top X)^{-1} \boldsymbol{x}_0^\top$$

Therefore,

$$\hat{Y}_0 \sim N(\boldsymbol{x}_0 \boldsymbol{\beta}, \sigma^2 \boldsymbol{x}_0 (X^\top X)^{-1} \boldsymbol{x}_0^\top)$$

$$\frac{\hat{Y}_0 - \boldsymbol{x}_0 \boldsymbol{\beta}}{\sigma \sqrt{\boldsymbol{x}_0 (X^\top X)^{-1} \boldsymbol{x}_0^\top}} \sim N(0,1)$$

$$\frac{\hat{Y}_0 - \boldsymbol{x}_0 \boldsymbol{\beta}}{\hat{\sigma} \sqrt{\boldsymbol{x}_0 (X^\top X)^{-1} \boldsymbol{x}_0^\top}} \sim t(n-(p+1)) = t(n-p-1)$$

A $(1-\alpha)$ confidence interval for the mean response given $\boldsymbol{x}_0$,

$$\hat{y}_0 \pm c\hat{\sigma} \sqrt{\boldsymbol{x}_0 (X^\top X)^{-1} \boldsymbol{x}_0^\top}$$

where c is the $1-\alpha/2$ quantile of $t(n-p-1)$.

Prediction error: $Y_0 - \hat{Y}_0$ which are independent since Y_0 is a random variable with variance σ^2 and $\hat{Y}_0$ is a function of $Y_1, \ldots, Y_n$. Therefore,

$$\mathsf{E}[Y_0 - \hat{Y}_0] = \boldsymbol{x}_0 \boldsymbol{\beta} - \boldsymbol{x}_0 \boldsymbol{\beta} = 0$$

$$\mathsf{Var}(Y_0 - \hat{Y}_0) = \mathsf{Var}(Y_0) + (-1)^2 \mathsf{Var}(\hat{Y}_0) = \sigma^2 + \sigma^2 (\boldsymbol{x}_0 (X^\top X)^{-1} \boldsymbol{x}_0^\top)$$

Therefore,

$$Y_0 - \hat{Y}_0 \sim N(0, \sigma^2 (1 + \boldsymbol{x}_0 (X^\top X)^{-1} \boldsymbol{x}_0^\top))$$

A $(1-\alpha)$ prediction interval for y_0

$$\hat{y}_0 \pm c\hat{\sigma} \sqrt{1 + \boldsymbol{x}_0 (X^\top X)^{-1} \boldsymbol{x}_0^\top}$$

where c is the $1-\alpha/2$ quantile of $t(n-p-1)$.

Intuition: prediction interval wider than CI for mean. Estimating an average is "easier" than an individual response.