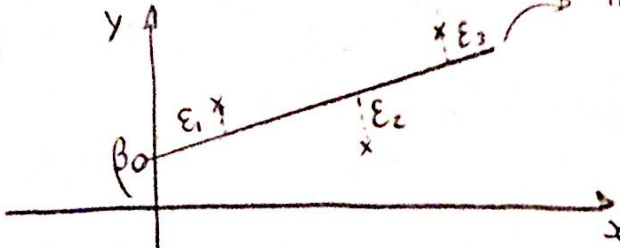
① Prediction in Linear Regression

Recall



True Regression line
$Y = \beta_0 + \beta_1 x$
x : data points

Simple Linear Regression model : $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \ldots, n$ (data points)
parameters $\beta_0, \beta_1, \sigma^2$
σ^2 → error variance (ε is assumed to be $\sim N(0, \sigma^2)$)

The fitted model: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

(the purpose of collecting data is to estimate and make inference about the parameters β_0, β_1 and σ^2).

Using LSE, we obtained equations for $\hat{\beta}_0$ and $\hat{\beta}_1$ as follows:

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} \quad \text{and} \quad \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

For $\hat{\sigma}^2$ we don't use LSE, but an intutive estimator of σ^2 is

$$MSE = \hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \sum (y_i)^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i \sim \frac{\sigma^2 \chi^2_{n-2}}{n-2}$$

The quantities $\varepsilon_i = y_i - \hat{y}_i \; (i = 1, \ldots, n)$ are called residuals

As we have proven before, all estimators in LR are unbiased, i.e.,

$$E[\hat{\beta}_0] = \beta_0, \quad E[\hat{\beta}_1] = \beta_1, \quad E[\hat{\sigma}^2] = \sigma^2$$

with variances

$$var(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right), \quad var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad cov(\hat{\beta}_0, \hat{\beta}_1) = \frac{\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sigma^2$$

→ Predicted, or fitted, values are the values of y predicted by LSE regression line ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$) obtained by plugging-in values of x.

For instance if my fitted model is $\hat{y} = \underbrace{1}_{\hat{\beta}_0} + \underbrace{0.5}_{\hat{\beta}_1} x$

The predicted value for $x = 1$ is $\hat{y} = 1 + 0.5(1) = 1.5$

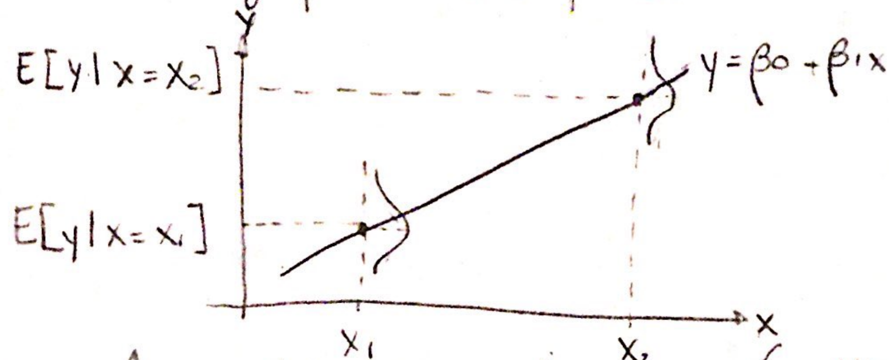
→ Mean response prediction / estimated average for a given value $x = x_0$.

Recall $E[\hat{y}] = E[\hat{\beta}_0 + \hat{\beta}_1 x] = \beta_0 + \beta_1 x$

What happens if we want to compute this expectation for $x = x_0$?

Intuition: imagine you're an uber driver and you're collecting data of miles traveled and amount of gas you put in the tank of your car ⇒ $gas = \hat{\beta}_0 + \hat{\beta}_1 (miles)$ is your fitted model. So, to calculate the mean gas volume for all trips (regardless of miles traveled) you can say: I anticipate that I will use $E[gas]$ gallons of fuel. However, in practice, you may be more interested in knowing $E[gas \mid miles = m]$ or, in words, given that I know I will travel m miles, how many gallows should I expect to use?

Here's the graphical interpretation:



⇒ for fixed x, the variable Y differs from its expected value by a random amount.

Hence: $\hat{y}_{x_0} = E[\hat{y} \mid X = x_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0$ (conditional expected value of Y given $X = x_0$)

and $var[\hat{y} \mid X = x_0] = var(\hat{\beta}_0 + \hat{\beta}_1 x_0) = var(\hat{\beta}_0) + x_0^2 \, var(\hat{\beta}_1) + 2x_0 \, cov(\hat{\beta}_0, \hat{\beta}_1)$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\Sigma(x_i - \bar{x})^2}\right) + \frac{x_0^2 \, \sigma^2}{\Sigma(x_i - \bar{x})^2} + 2x_0 \frac{\bar{x}\,\sigma^2}{\Sigma(x_i - \bar{x})^2}$$

$$= \sigma^2\left(\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

A $100(1-\alpha)\%$ CI for the estimated mean response when $x = x_0$ is:

$$\hat{y}_{x_0} \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2} \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}}$$

For MLR: $x^T\hat{\beta} \pm t_{n-k-1,\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2 x^T (X^T X)^{-1} x}$

→ Prediction interval for a single future observation

We now consider the prediction of a new or future observation y corresponding to $x = x_0$. Therefore, the predicted value $\hat{y}_{x_0}$ is still obtained by substituting $x = x_0$ in the fitted regression model ($\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$) and is the same as the estimated mean response in the previous page.

However, by predicting an individual outcome, the variance σ^2 of an individual observation should be added to the variance of $\hat{y}_{x_0}$, since the future value actually observed will fluctuate around the mean response value.

Recall the graphical interpretation of the previous page? The normal shaped curves around the values of y_{x_1} and y_{x_2} represent variability around mean response value.

By adding σ^2 (variance of predicted observation) to $var[\hat{y} \mid X = x_0]$:

$$var(\hat{y}_{pred,x_0}) = Var(\hat{y}_{pred} \mid X = x_0) = var(\hat{\beta}_0 + \hat{\beta}_1 x_0) + \underbrace{var(x_0)}_{\sigma^2}$$

$$= \sigma^2\left(1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

a $100(1-\alpha)\%$ prediction interval for a single response is:

$$\hat{y}_{pred,x_0} \pm t_{n-2,\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2} \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

For MLR: $$x^T\hat{\beta} \pm t_{n-k-1,\frac{\alpha}{2}} \sqrt{\hat{\sigma}^2(1 + x^T(X^TX)^{-1}x)}$$

② ANOVA

Another way of judging the explanatory power of covariates (other than hypothesis testing) is to build an ANOVA table. This table splits the variation in the data. For Simple Linear Regression, the ANOVA Table is:

Source	Degrees of freedom	Sum of Squares	Mean Sum of Squares	F-stats	p-value
Regression (RegSS)	1	$\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$	$\hat{\beta}_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$	$\frac{\hat{\beta}_1^2 \sum_{i=1}^n (x_i-\bar{x})^2}{\sum_{i=1}^n (y_i-\hat{y}_i)^2/(n-2)}$	$P(F_{1,n-2} > F_{obs})$
Residual (SSE)	$n-2$	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$\sum_{i=1}^n \underbrace{(y_i-\hat{y}_i)^2}_{e_i^2}/(n-2)$		
Total (CTSS)	$n-1$	$\sum_{i=1}^n (y_i - \bar{y})^2$			

$\Rightarrow MSE = \hat{\sigma}^2$

$H_0: \beta_1 = 0$

Where $\underset{\text{between group variability}}{\text{RSS}}$ represents the variation explained by the model and $\underset{\text{within group variability}}{\text{SSE}}$ represents the residual variation.

The following holds: $CTSS = RegSS + SSE$

and we define

$$R^2 = \frac{RegSS}{CTSS} = 1 - \frac{SSE}{CTSS}$$

{ R^2 measures the "proportion of total variation explained by the fitted regression model"

For General Linear Model:

Source	Df's	SS	MSS	F-stats	p-value
Regression (RegSS)	K	$\hat{\beta}^T X^T X \hat{\beta} - N\bar{y}^2$	$RegSS / K$	$\frac{RegSS/K}{SSE/(n-K-1)}$	$P(F_{K,N-K-1} > F_{obs})$
Residual (SSE)	$n-K-1$	$(y - X\hat{\beta})^T (y - X\hat{\beta})$	$SSE/(n-K-1)$		
Total (CTSS)	$n-1$	$y^T y - N\bar{y}^2$	$CTSS/(n-1)$		

$H_0: \beta_1 = \beta_2 = \dots = \beta_p$

Multiple Regression Example:

After running a multiple regression with 2 covariates (X1 and X2) in R, the following table was obtained (n=20).

Predictor	Coef.	SE coef.	T
constant	3.324	3.111	1.07
X1	3.7681	0.6142	$\boxed{}$
X2	5.0796	0.6655	$\boxed{}$

① Compute t-value corresponding to X1 and X2.

Answer: $t_{X1} = \frac{3.7681}{0.6142} = 6.13$ $\quad t_{X2} = \frac{5.0796}{0.6655} = 7.63$

② Test whether the predictor variables significantly affect Y at level 0.05.

$n = 20$ so $t_{n-\overset{2}{p}-1} = t_{20-2-1} = t_{17}$

$t_{17, 1-\frac{0.05}{2}} = t_{17, 0.975} = 2.109816$ (in R qt(0.975, 17))

since both t_{X1} and t_{X2} are $> t_{CRITICAL}$ ($= 2.109816$), both t_{X1} and t_{X2} are significant at 0.05 level

③ Using the model, estimate the value of Y at $X_1 = 5$ and $X_2 = 16$.

$$\hat{Y} = 3.324 + 3.7681(\underset{5}{X1}) + 5.0796(\underset{16}{X2}) = 103.44$$

... Coming next class (10/06) { • CI for estimated average
• Prediction interval

Continuing the example

(4) Suppose $(X^TX)^{-1}$ is given below

$$(X^TX)^{-1} = \begin{bmatrix} 0.307 & -0.033 & 0.015 \\ -0.033 & 0.012 & -0.012 \\ 0.015 & -0.012 & 0.014 \end{bmatrix}$$

using $X_1 = 5$ and $X_2 = 16$, Find 95% CI interval for the estimated average.

From (3) $\hat{y} = 103.44$

To get $\hat{\sigma}^2$ we can do $\quad var(\hat{\beta}_1) = \hat{\sigma}^2((X^TX)^{-1}_{22}) = 0.012\,\hat{\sigma}^2$

$$\underbrace{(se(\hat{\beta}_1))^2}_{} (0.6142)^2 = 0.012\,\hat{\sigma}^2 \quad \Rightarrow \hat{\sigma}^2 \approx 31.44$$

$X_{new} = [1, 5, 16]^T \qquad X^T_{new}(X^TX)^{-1}X_{new} = 2.421$

$$\hat{y}_{new} \pm \underbrace{t_{17,0.025}}_{2.110}\,\hat{\sigma}\sqrt{X^T_{new}(X^TX)^{-1}X_{new}} = [\,85.03\,,\,121.8\,]$$

(5) Obtain the 95% prediction interval for a singular future observation $\hat{y}_{pred,new}$ using same X as in (4).

$$\hat{y}_{pred,new} \pm t_{17,0.025}\,\hat{\sigma}\sqrt{1 + X^T_{new}(X^TX)^{-1}X_{new}} = [81.56\,,\,125.3]$$

↳ this is wider! Do you remember why?