Inference

① Sampling distribution of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$

Recall that

$E(\hat{\beta}_1) = \beta_1$, $Var(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$,

$E(\hat{\beta}_0) = \beta_0$, $Var(\hat{\beta}_0) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}_1^2}{S_{xx}}\right)$

and $E(\hat{\sigma}^2) = \sigma^2$

These results rely on the SLR assumption

$E(\varepsilon \mid X_1 = x_1) = 0$, $Var(\varepsilon \mid X_1 = x_1) = \sigma^2$

⇒ Additional assumption: Gaussian-Noise SLR

⇒ conditional Gaussian distributed Y.

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

$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}_1^2}{S_{xx}}\right)\right)$

$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$ ($p=2$ for SLR)

1.1 Gaussian-Noise Simple Linear Regression Model

Define Gaussian-Noise SLR as

same:

① The distribution of X is arbitrary (and perhaps X is even non-random)

② $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

If $X_1 = x_1$, then $Y = \beta_0 + \beta_1 x_1 + \varepsilon$ for some coefficients β_0, β_1 and random noise ε

$E(\varepsilon)=0$, $Var(\varepsilon)=\sigma^2$

③ $\varepsilon \sim N(0, 1)$

④ ε is independent of X_1 and independent across observations.

ε is uncorrelated with X_1

Linear correlation

Those having 0 correlation might dependent non-linearly

Stronger assumption

1.2 Sampling distribution of $(\hat{\beta}-\beta)/se(\hat{\beta})$ and $(\hat{\beta}-\beta)/ese(\hat{\beta})$

Since $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$ $\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n}+\frac{\bar{x}_1^2}{S_{xx}}))$

$\Rightarrow \frac{\hat{\beta}_1-\beta_1}{se(\hat{\beta}_1)} \sim N(0,1)$ (σ unknown) $\frac{\hat{\beta}_0-\beta_0}{se(\hat{\beta}_0)} \sim N(0,1)$ (σ unknown)

Replace $se(\hat{\beta}_1)$ and $se(\hat{\beta}_0)$ with $ese(\hat{\beta}_1)$ and $ese(\hat{\beta}_0)$

$\Rightarrow T_1 = \frac{\hat{\beta}_1-\beta_1}{ese(\hat{\beta}_1)} \sim t_{n-p}$ $T_0 = \frac{\hat{\beta}_0-\beta_0}{ese(\hat{\beta}_0)} \sim t_{n-p}$

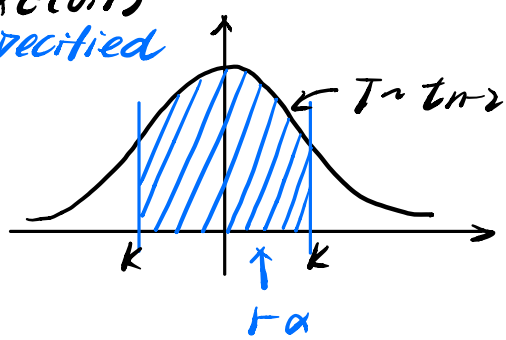
where $p=2$ for GN-SLR

We can compute $ese(\hat{\beta}_1) = \sqrt{\frac{MS_{Res}}{S_{xx}}}$ and $ese(\hat{\beta}_0) = \sqrt{MS_{Res}(\frac{1}{n}+\frac{\bar{x}_1^2}{S_{xx}})}$ from the data without knowing σ^2.

② Confidence Interval

For β_1. Suppose that f_T is the density of a $T \sim t_{n-2}$ distribution, define $k \equiv t_{\frac{\alpha}{2}, n-2} > 0$

such that $P(-k \le T \le k) = 1-\alpha$ $\alpha \in (0,1)$ (user specified)



A $1-\alpha$ level confidence interval for β_1 (not $\hat{\beta}_1$) is

$CI(\beta_1) = [\hat{\beta}_1 - k \cdot ese(\hat{\beta}_1), \hat{\beta}_1 + k \cdot ese(\hat{\beta}_1)]$

$P(\beta_1 \in CI(\beta_1))$

$= P(\hat{\beta}_1 - k \cdot ese(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + k \cdot ese(\hat{\beta}_1))$

$= P(-k \le \frac{\hat{\beta}_1-\beta_1}{ese(\hat{\beta}_1)} \le k)$ ($\frac{\hat{\beta}_1-\beta_1}{ese(\hat{\beta}_1)} = T_1$)

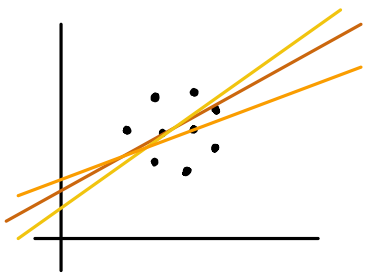
$= 1-\alpha.$

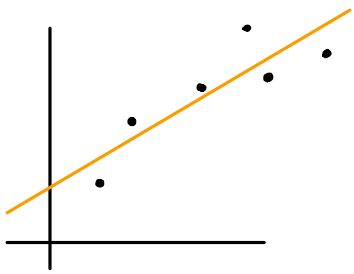
	SLR	GN-SLR	
Estimation $\hat{\beta}_1$, $\text{esc}(\hat{\beta}_1)$, $\hat{\sigma}^2$	✓	✓	not necessary
unbiaseness $E(\hat{\beta}_1)$ variance $\text{Var}(\hat{\beta}_1)$	✓	✓	not necessary
confidence interval		✓	
hypothesis testing (t-test)		✓	
F-test		✓	
Prediction interval		✓	

The upper and lower bound for $CI(\beta_1)$ are random, since their formula involve the estimator $\hat{\beta}_1$, which could be potentially random if $\hat{\beta}_1$ is computed using a different set of samples.

[Interpretation]

- The interval $CI(\beta_1)$ traps β_1 with probability $1-\alpha$.
- β_1 is non-random
- The width of the CI is $2 \cdot k \, ese(\hat{\beta}_1)$
  - As α shrinks, the interval widens
  - As n grows, the interval shrinks.
  - As σ^2 increases, the interval widens.
  - As S_{xx} grows, the interval shrinks.





A $1-\alpha$ $CI(\beta_0)$ is

$$CI(\beta_0) = [\hat{\beta}_0 - k \cdot ese(\hat{\beta}_0), \hat{\beta}_0 + k \cdot ese(\hat{\beta}_0)]$$

k remains the same as long as the distribution and sample size don't change.

② Hypothesis Testing

two sided test $\begin{cases} H_0: \beta_1 = c \\ H_1: \beta_1 \neq c \end{cases}$

Rule: We reject H_0 in hypothesis test with significant level α, if $100(1-\alpha)\%$ confidence interval $CI(\beta_1)$ doesn't cover c.

Why: Let's set $\alpha = 0.01$ for example.
Because if H_0 is true, then highly likely (with probability $1-\alpha = 0.99$) $CI(\beta_1)$ should trap $\beta_1 = c$.

- either in fact $\beta_1 = c$ (H_0 is true), and therefore you just (by chance) observe a very rare event that $CI(\beta_1)$ doesn't trap $\beta_1 = c$ (with probability $\alpha = 0.01$)
- or just because $\beta_1 \neq c$, and H_0 should be rejected.

If we reject H_0 when $CI(\beta_1)$ doesn't cover c i.e. $c \notin CI(\beta_1)$

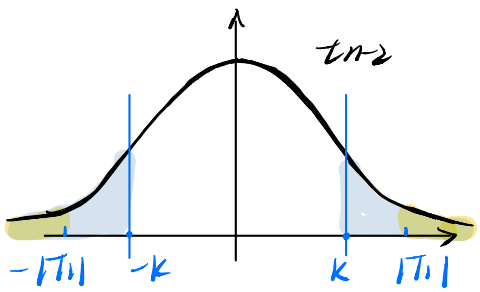
$\Leftrightarrow$ $c \geq \hat{\beta}_1 + k \cdot ese(\hat{\beta}_1)$ or $c \leq \hat{\beta}_1 - k \cdot ese(\hat{\beta}_1)$

$\Leftrightarrow$ $\underbrace{\frac{\hat{\beta}_1 - c}{ese(\hat{\beta}_1)}}_{T_1} \leq -k$ or $\underbrace{\frac{\hat{\beta}_1 + c}{ese(\hat{\beta}_1)}}_{T_1} \geq k$

$\Leftrightarrow$ $|T_1| \geq k \equiv t_{\frac{\alpha}{2}, n-2}$ $\quad T_1 = \frac{\hat{\beta}_1 - c}{ese(\hat{\beta}_1)}$ t test statistic

Wald Test or t-test

$\Leftrightarrow$ $Pr(|T| > |T_1|) < \alpha$
p-value



For β_0

$\begin{cases} H_0: \beta_0 = c \\ H_1: \beta_0 \neq c \end{cases}$

We reject H_0 if $|T_0| \geq k \equiv t_{\frac{\alpha}{2}, n-2}$ $\quad T_0 = \frac{\hat{\beta}_0 - c}{ese(\hat{\beta}_0)}$

3.1 Testing Significance of Regression

Failing to reject $H_0: \beta_1 = 0$ implies that there is no linear relationship between x_1 and Y.

- either that x_1 is of little value in explaining the variation in Y.
- or that the true relationship between x and Y is not linear.

We reject H_0 with significant level α if the $100(1-\alpha)\%$ confidence interval $CI(\beta_1)$ doesn't cover 0.

The corresponding test statistics is

$$T_1 = \frac{\hat{\beta}_1 - 0}{ese(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{ese(\hat{\beta}_1)}$$

We reject $H_0: \beta_1 = 0$ if $|T_1| > k \equiv t_{\frac{\alpha}{2}, n-1}$

3.2 Statistical Significance

- What α should we use? Conventionally set $\alpha = 0.05$. If you're very conservative and cannot afford false rejection, then you should use smaller α.
- If we test the hypothesis that $H_0: \beta_1 = 0$ and reject, we say that the difference between β_1 and 0 is statistically significant.
- statistical significance ≠ importance of magnitude.

  When H_0 is attained, it should be interpreted as

  "the effect of β_1 is statistically undetectable." or

  "β_1 is statistically undistinguishable from 0."

No strong evidence to show.....

e.g. A common fallacy is

> "I tested wether $H_0: \beta_0 = 0$ or not, and I retain H_0. Therefore, β_1 is insignificant, and it is unimportant and you can ignore it."

Because we retain $H_0: \beta_1 = 0$ when $\left| \frac{\hat{\beta}_1}{\text{esel}(\hat{\beta}_1)} \right| < k$, there are two cases that might result in a retain of H_0:

- $\hat{\beta}_1$ is very close to zero, which implies β_1 may be as well as zero, thus unimportant
- $\text{esel}(\hat{\beta}_1)$ is so large that we can't tell anything about β_1 with any confidence $\Rightarrow$ β_1 may be either large or small.

④ Analysis of Variance (ANOVA)

The distance from any point y_i in a collection of data to $\bar{y}$, is the direction written as $y_i - \bar{y}$.

The ANOVA is based on a partition of total variability in y.

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

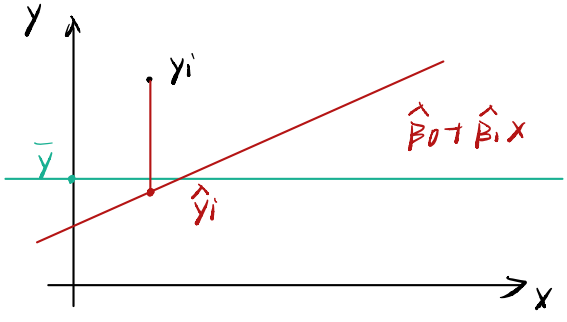
$$\Rightarrow \quad \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

SS_T SS_{Res} SS_R

Total sum of squares. It measures the "total variance" in Y that can be decomposed to two components.

Residual sum of squares It measures amount of "residual variance" in Y. It is an aggregate measure of mis-fitted regression line.

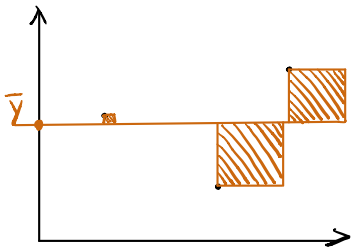
Regression sum of squares It measures amount of "systematic variation" in Y due to the Y~X linear relationship.



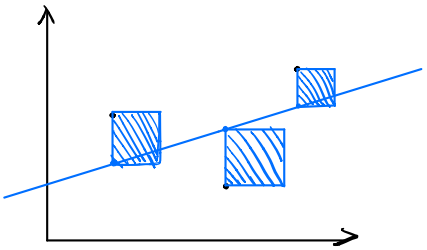
$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

cannot be explained by regression

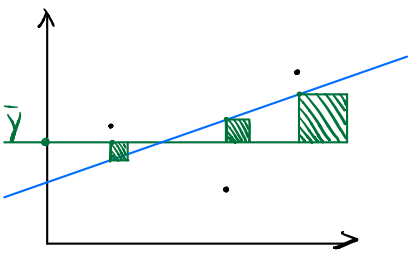
variation explainable by regression



$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



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



$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - \bar{y})^2 &= \sum_{i=1}^{n} (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} ((\underbrace{y_i - \hat{y}_i}_{e_i}) + (\hat{y}_i - \bar{y}))^2 \\
&= \sum_{i=1}^{n} ((\hat{y}_i - \bar{y})^2 + 2e_i(\hat{y}_i - \bar{y}) + e_i^2) \\
&= \underbrace{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n} e_i^2}_{SSRes} + 2\sum_{i=1}^{n} e_i(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} - \bar{y}) \\
&= SSR + SSRes + 2(\hat{\beta}_0 - \bar{y}) \sum_{i=1}^{n} e_i + 2\hat{\beta}_1 \underbrace{\sum_{i=1}^{n} e_i x_{i1}}_{0}
\end{aligned}$$

⑤ F-test (intercept vs. whole model)

The idea is to compare two models

$H_0: Y = \beta_0 + \varepsilon$

$H_1: Y = \beta_0 + \beta_1 X_1 + \varepsilon$

or equivalent to

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0.$

If we fit the first model, the least squares estimator is $\hat{\beta}_0 = \bar{y}$

The idea is to create a statistic that measures how much better the second model is than the first model.
Under GN-SLR, and under $H_0: \beta_1 = 0$, we show that

$$\frac{SS_T}{\sigma^2} \sim \chi^2_{n-1} \qquad df = n-1$$

$$\frac{SS_{RES}}{\sigma^2} \sim \chi^2_{n-p} \qquad df_{RES} = n-p = n-2$$

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{p-1} \qquad df_R = p-1 = 2-1 = 1$$

for GN-SLR, $p = 2$

The degree of freedom has an additive property

$$df_T = df_R + df_{RES}$$

Consider now the F-test statistic.

$$F_0 = \frac{SS_R / df_R}{SS_{RES} / df_{RES}} = \frac{SS_R / (p-1)}{SS_{RES} / (n-p)} = \frac{MS_R}{MS_{RES}}$$

under $H_0: \beta_1 = 0$

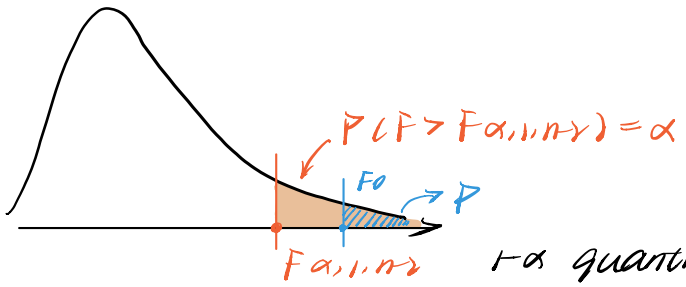
The statistic F_0 follows F distribution, with degrees of freedom $p-1$ and $n-p$

$$F_0 \sim F_{p-1, n-p} = F_{1, n-2}$$

We reject H_0 if F_0 is "large enough"

We set significance level α 0.01 or 0.05 and reject H_0 if $F_0 > F_{0\alpha, p-1, n-p}$

$\Leftrightarrow$ We reject H_0 if p-value $< \alpha$ $\quad Pr(F > F_0) < \alpha = Pr(F > F_{\alpha, 1, n-2})$



$1-\alpha$ quantile of the F-distribution with df 1 and $n-2$

Source	SS	df	MS	F
Regression	SS_R	$p-1$	$MS_R = \frac{SS_R}{p-1}$	$\frac{SS_R/(p-1)}{SS_{Res}/(n-p)}$
Residual	SS_{Res}	$n-p$	$MS_{Res} = \frac{SS_{Res}}{n-p}$	
Total	SS_T	$n-1$		

5.1 Equivalence of F-test and t-test for SLR.

$$(T_1)^2 = \left(\frac{\hat{\beta}_1}{ese(\hat{\beta}_1)}\right)^2 = F_0$$

The p-value for F-test is exactly the same as the p-value for t-test

$$F > F_0 \iff T^2 > (T_1)^2$$

Thus, $P(F > F_0) = P(T > T_1)$

Therefore, the test result based on F and t are the same.

Now we show that $(T_1)^2 = F_0$, since by

$$\begin{aligned} SS_R &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} [(\hat{\beta}_0 + \hat{\beta}_1 x_{i1}) - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1)]^2 \\ &= \hat{\beta}_1^2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 \\ &= \hat{\beta}_1^2 S_{xx} \end{aligned}$$

and $ese(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$

$$F = \frac{SS_R/(p-1)}{SS_{Res}/(n-p)} = \frac{SS_R/1}{SS_{Res}/(n-p)} = \frac{SS_R}{\hat{\sigma}^2} = \frac{\hat{\beta}_1^2 S_{xx}}{ese(\hat{\beta}_1)^2 S_{xx}} = \left[\frac{\hat{\beta}_1}{ese(\hat{\beta}_1)}\right]^2 = T_1^2$$

5.2 What the F-test really tests?

1. If retain $H_0: \beta_1 = 0$ – do not find any significant share of variance associated with the regression.
   (a) The intercept-only model is correct.
   (b) $\beta_1 \neq 0$, but the test doesn't have enough power to detect it $\frac{\hat{\beta}_1}{ese(\hat{\beta}_1)}$ small
   (c) The relationship is non-linear, but the best linear approximation has zero slope.

2. If reject H_0 – doesn't mean the SLR model is right, only that the non-zero slope linear model predict better than the intercept-only model.

⑥ Prediction and Predictive Inference

6.1 Prediction.

The conditional mean of Y is

$$\mu = m(x_1) = E(Y|X_1 = x_1) = \beta_0 + \beta_1 x_1$$

while our estimate of the conditional mean is:

$$\hat{\mu} = \hat{m}(x_1) = \hat{\beta}_0 + \hat{\beta}_1 x_1.$$

NOTE: $m(x_1)$ is deterministic, but unknown.
$\hat{m}(x_1)$ is a function of data, random, known. It inherits the randomness from $\hat{\beta}_0$ and $\hat{\beta}_1$, which in turn inherits from Y.

We can show that

$$E(\hat{m}(x_1)) = \beta_0 + \beta_1 x_1 = m(x_1)$$

$$Var(\hat{m}(x_1)) = \sigma^2 \left(\frac{1}{n} + \frac{(x_1 - \bar{x}_1)^2}{S_{xx}}\right)$$

Under the GN-SLR assumption, $\hat{m}(x_1)$ is Gaussianal.

$$\hat{m}(x_1) \sim N\left(m(x), \sigma^2\left(\frac{1}{n} + \frac{(x_1-\bar{x}_1)^2}{S_{xx}}\right)\right)$$

A $100(1-\alpha)$ percent CI on the mean response, at the point $x_1 = x_{01}$

conditional mean of Y.

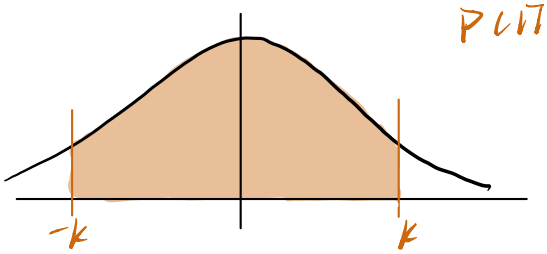
$$CI(m(x_{01})) = \left[\hat{m}(x_{01}) \mp k\sqrt{\frac{MS_{Res}}{\hat{\sigma}^2}\left(\frac{1}{n} + \frac{(x_{01}-\bar{x}_1)^2}{S_{xx}}\right)}\right]$$

$$= \left[\hat{m}(x_{01}) \pm k \cdot \frac{se(\hat{m}(x_{01}))}{SS}\right]$$

$\sqrt{var(\hat{m}(x_{01}))}$ with σ^2 replaced by $\hat{\sigma}^2$.

where $k = t_{\frac{\alpha}{2}, n-2}$.

$P(|T| < k) = \alpha$



A $100(1-\alpha)$ percent CI on the future observation is

$$CI(y) = \left[\hat{m}(x_{01}) \mp k \cdot \sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_{01}-\bar{x}_1)^2}{S_{xx}}\right)}\right]$$

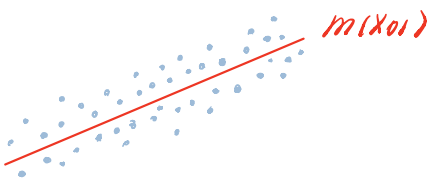
We should distinguish mean response $CI(m(x_{01}))$ and new observation $CI(y)$

- A $100(1-\alpha)\%$ CI on the mean response is an interval $[l, u]$ where

$$P(l \le m(x_1) \le u \mid x_1 = x_{01}) = 1-\alpha$$

recall: $Y = \underbrace{\beta_0 + \beta_1 x_1}_{m(x_1)} + \varepsilon$

$E(Y \mid X_1 = x_1) = m(x_1)$



- A $100(1-\alpha)\%$ CI on the new observation Y is an interval $[l', u']$ where

$$P(l' \le \underbrace{Y_{x01}}_{Y_{x01} = \beta_0 + \beta_1 x_1 + \varepsilon} \le u' \mid x_1 = x_{01}) = 1-\alpha$$

Proof

Mean Response

We predict on $x_1 = x_{01}$

$$\hat{m}(x_{01}) = \hat{\beta}_0 + \hat{\beta}_1 x_{01}$$
$$= (\bar{y} - \hat{\beta}_1 x_{01}) + \hat{\beta}_1 x_{01}$$
$$= \bar{y} + (x_{01} - \bar{x}_1)\hat{\beta}_1$$

using $\hat{\beta}_1 = \sum_{i=1}^{n} c_i y_i$, $c_i = \frac{x_{i1} - \bar{x}_1}{S_{xx}}$ $\quad \sum_{i=1}^{n} c_i = 0$

Now plug-in $\hat{\beta}_1$

$$\hat{m}(x_{01}) = \frac{1}{n}\sum_{i=1}^{n} y_i + (x_{01} - \bar{x}_1)\sum_{i=1}^{n} c_i y_i$$

$$E(\hat{m}(x_{01})) = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i + (x_{01} - \bar{x}_1)\sum_{i=1}^{n} c_i y_i\right]$$

① $E\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_{i1} + \varepsilon_i)\right] = \beta_0 + \beta_1 \bar{x}_1$

② $E\left[(x_{01} - \bar{x}_1)\sum_{i=1}^{n} c_i y_i\right] = (x_{01} - \bar{x}_1)\beta_1$ $\qquad E(\hat{\beta}_1) = \beta_1$

$$\Rightarrow E(\hat{m}(x_{01})) = \beta_0 + \beta_1 \bar{x}_1 + (x_{01} - \bar{x}_1)\beta_1 = \beta_0 + \beta_1 x_{01}$$

$$\mathrm{Var}(\hat{m}(x_{01})) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i + (x_{01} - \bar{x}_1)\left(\sum_{i=1}^{n} c_i y_i\right)\right)$$
$$= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\left(1 + n(x_{01} - \bar{x}_1)c_i\right) y_i\right)$$
$$= \frac{1}{n^2}\sum_{i=1}^{n}\left(1 + n(x_{01} - \bar{x}_1)c_i\right)^2 \mathrm{Var}(y_i)$$

$y_i, y_j \mid x_1 = x_{01}$ independent

$$= \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\left(1 + 2n(x_{01} - \bar{x}_1)c_i + n^2(x_0 - \bar{x}_1)^2 c_i^2\right)$$
$$= \frac{\sigma^2}{n^2}\left(n + 0 + n^2(x_0 - \bar{x}_1)^2 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}{S_{xx}^2}\right)$$
$$= \frac{\sigma^2}{n^2}\left(n + 0 + \frac{n^2(x_0 - \bar{x}_1)^2}{S_{xx}}\right)$$
$$= \sigma^2\left(\frac{1}{n} + \frac{(\bar{x}_1 - x_{01})^2}{S_{xx}}\right)$$

cannot eliminate.

Note: The variance of $m(x_{01})$ grows as σ^2 goes.
The larger the n, the smaller the variance
The further our prediction point x_{01} is from the center of data, the bigger the variance.

Thus, under GN-SLR, $\hat{m}(x_{01})$ is gaussian

$$\hat{m}(x_{01}) \sim N\left(E(\hat{m}(x_{01})), \operatorname{var}(\hat{m}(x_{01}))\right)$$
$$= N\left(m(x_{01}), \sigma^2\left(\frac{1}{n} + \frac{(x_{01}-\bar{x}_1)^2}{S_{xx}}\right)\right)$$

$$\frac{\hat{m}(x_{01}) - m(x_{01})}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{(x_{01}-\bar{x}_1)^2}{S_{xx}}\right)}} \sim N(0,1)$$

$$\frac{\hat{m}(x_{01}) - m(x_{01})}{\operatorname{ese}(\hat{m}(x_{01}))} = \frac{\hat{m}(x_{01}) - m(x_{01})}{\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_{01}-\bar{x}_1)^2}{S_{xx}}\right)}} \sim t_{n-2}$$

Thus, a $100(1-\alpha)\%$ CI on the mean response at point $x_1 = x_{01}$ is

$$CI(m(x_1)) = \left[\hat{m}(x_{01}) \mp k\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_{01}-\bar{x}_1)^2}{S_{xx}}\right)}\right]$$

Future Observation

using GN-SLR model, we know that at $x_1 = x_{01}$,
the future observation of Y_0 is

$$Y_0 = m(x_{01}) + \varepsilon = \beta_0 + \beta_1 x_{01} + \varepsilon$$

Now we could construct a confidence interval for Y_0
or prediction interval

We call it the prediction interval for the future observation Y_0.
(which is predicted at $x_1 = x_{01}$)

$$PI(y_0) = \left[\hat{Y}_0 \mp k \cdot \operatorname{ese}(\hat{Y}_0)\right]$$
$$= \left[\hat{m}(x_{01}) \mp k \cdot \sqrt{MS_{Res}\left(1 + \frac{1}{n} + \frac{(x_{01}-\bar{x}_1)^2}{S_{xx}}\right)}\right]$$

wider than $CI(m(x_{01}))$

Since $Y_0 \mid X_1 = x_{01} = m(x_{01}) + \varepsilon \sim N(m(x_{01}), \sigma^2)$

The variance of Y_0 is $\text{Var}(Y_0) = \text{Var}(m(x_{01}) + \varepsilon)$

$$= \text{Var}(m(x_{01})) + \text{Var}(\varepsilon)$$
$$= \text{Var}(\beta_0 + \beta_1 x_{01}) + \sigma^2$$
$$= \sigma^2.$$

In the formula, σ^2, β_0, β_1 are unknown, replace them with $\hat{\sigma}^2$, $\hat{\beta}_0$, $\hat{\beta}_1$

$$\hat{\text{Var}}(Y_0) = \hat{\text{Var}}(\hat{m}(x_{01}) + \varepsilon)$$
$$= \hat{\text{Var}}(\hat{\beta}_0 + \hat{\beta}_1 x_{01}) + \hat{\text{Var}}(\varepsilon)$$
$$= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_{01} - \bar{x}_1)^2}{S_{xx}}\right) + \hat{\sigma}^2$$
$$= \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_{01} - \bar{x}_1)^2}{S_{xx}}\right)$$

Therefore, the estimated standard error of Y_0 is

$$\text{ese}(Y_0) = \sqrt{\hat{\text{Var}}(Y_0)} = \sqrt{1 + \frac{1}{n} + \frac{(x_{01} - \bar{x}_1)^2}{S_{xx}}}$$

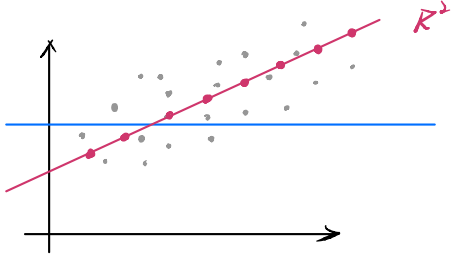
① R^2 (rubbish)

If we want to get a global measure of how well X_1 predict Y. you may consider the proportion of variation explained the regression model.

$$R^2 = \frac{SSR}{SST} = \frac{SST - SS_{Res}}{SST} = 1 - \frac{SS_{Res}}{SST}$$

Because $SSR = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx}$

- R^2 will be 0 if $\hat{\beta}_1 = 0$
- R^2 will be 1 if all the residuals are zero $SS_{Res} = 0$.



$R^2 = 1$ best case scenario

$\hat{\beta}_1 = 0 \Rightarrow R^2 = 0$ worst case scenario

Interpretation of R^2.

Distinguish R^2 from t-test (p-value)

① Low R^2 and low p-value (p-value $< \alpha$)

The model doesn't explain much of variation of the data but the model is significant.

(Including X_1 into the model is necessary but not sufficient).

② Low R^2 and high p-value

The model doesn't explain much of variation of the data and X_1 is not useful <worst case>

③ High R^2 and low p-value

Model sufficient, X_1 is useful <best case>

④ High R^2 and high p-value

Adjusted R^2

We have $E(SS_{Res}) = \sigma^2(n-p)$ (# variables)

If we increase the model complexity $p \to n$

$SS_{Res} \to 0 \quad R^2 = 1 - \frac{SS_{Res}}{SS_T} \to 1$

R^2 doesn't account for model "complexity".

adjusted R^2

$$R^2_{adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}$$

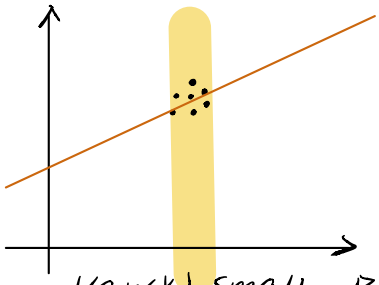
Note:

- When $p=1$, $R^2_{adj} = R^2$
- Adjusted R^2 is better than R^2 in term of model selection.

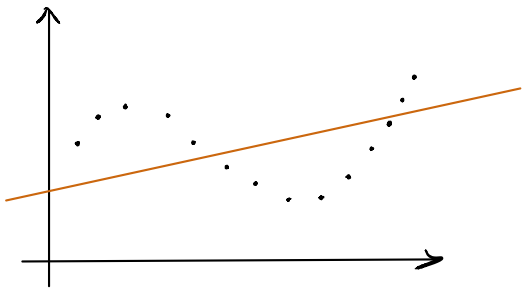
Limitation of R^2 (or R^2_{adj})

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSRes} = \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\beta}_1^2 S_{xx} + SSRes} = \frac{\hat{\beta}_1^2 \frac{S_{xx}}{n}}{\hat{\beta}_1^2 \frac{S_{xx}}{n} + \frac{SSRes}{n}} \xrightarrow{n\to\infty} \frac{\beta_1^2 Var(X)}{\beta_1^2 Var(X) + \sigma^2}$$

$$\hat{\sigma}^2 = \frac{SSRes}{n} \xrightarrow{n\to\infty} \sigma^2$$



$Var(X)$ small, R^2 will be small
model is correct,
but $R^2 \to 0$



$Var(X)$ large, $R^2 \to 1$
model is wrong, but $R^2 \approx 1$.

Note:

(a) We can manipulate the data that
$R^2 = 0$ but the model is correct
$R^2 = 1$ but the model is wrong

(b) R^2 can be compared when different models are fit to the same dataset with the same, untransformed response variable.
R^2 cannot be compared across different datasets.