MATH 423

Applied Regression

lec 1 (Sep 8) · Optimal Prediction

Optimal prediction for a single r.v. :

Say we have an r.v. Y, whose distribution is unknown.

Theorem : The optimal prediction m^* minimizes the expected mean squared error,

$$m^* = \underset{m}{\operatorname{argmin}}\ \mathbb{E}((Y-m)^2)$$

The optimal predictor of Y is $\mathbb{E}(Y)$, i.e. $m^* = \mathbb{E}(Y)$:

→ Proof :

$$\frac{\partial\, MSE}{\partial m} = \frac{\partial\ \mathbb{E}((Y-m)^2)}{\partial m} \qquad \text{recall that } Var(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

$$= \frac{\partial[(\mathbb{E}(Y-m))^2 + Var(Y-m)]}{\partial m}$$

$$= \frac{\partial}{\partial m}(\mathbb{E}(Y) - m)^2 + \frac{\partial}{\partial m} var(Y)$$

$$= \frac{\partial}{\partial m}(\mathbb{E}(Y)^2 - 2m\,\mathbb{E}(Y) + m^2)$$

$$= -2\,\mathbb{E}(Y) + 2m$$

$$\frac{\partial\, MSE}{\partial m} := 0 \Rightarrow m^* = \mathbb{E}(Y) \quad \blacksquare$$

In other words, the best 1 number guess we could make for Y is just its expected value.

What if we used something else other than MSE to measure how good the pred. is?

e.g. mean absolute deviation (MAD)

$$MAD = \mathbb{E}(|Y-m|)$$

$$\Rightarrow m^* = \text{median}(Y)$$

The median is more stable than the mean, especially if data has a lot of outliers. But solving for the m^* is more difficult as this includes taking the derivative of the absolute value → ???

Optimal prediction of an r.v. from other variables

- input variables $\xrightarrow{\text{model}}$ output variables
- input vars (X) are vars that are correlated with output, aka features, predictors
- output vars (Y) measure the outcome of interest, aka dep. var.

Note that predictive models do not show causation, only correlation!

Say we have 2 r.v.s, X, Y. The joint distribution of (X,Y) is known. We would like to predict Y using X.

Theorem: The optimal prediction function $m^*(\cdot)$ minimizes the expected MSE :

$$m^*(\cdot) = \underset{m(\cdot)}{\operatorname{argmin}}\ \mathbb{E}_{X,Y}[(Y - m(X))^2]$$

The optimal predictor of Y is $\mathbb{E}(Y|X=x)$ i.e. $m^*(x) = \mathbb{E}(Y|X=x)$.

→ Proof: Denote $\mu(x) = \mathbb{E}_{Y|X}(Y|X=x)$

We want to prove that $m^*(x) = \mu(x)$.

$$\begin{aligned}\mathbb{E}_{X,Y}[(Y-m(X))^2] &= \mathbb{E}_{X,Y}[(Y-\mu(X)+\mu(X)-m(X))^2] \\ &= \mathbb{E}_{X,Y}[(Y-\mu(X))^2] + 2\,\mathbb{E}_{X,Y}[(Y-\mu(X))(\mu(X)-m(X))] \\ &\quad + \mathbb{E}_{X,Y}[(\mu(X)-m(X))^2] \\ &= \mathbb{E}_{X,Y}[(Y-\mu(X))^2] + \mathbb{E}_X[(\mu(X)-m(X))^2]\end{aligned}$$

Notice that $\mathbb{E}_{X,Y}[(Y-\mu(X))(\mu(X)-m(X))] = 0$:

$$\begin{aligned}\mathbb{E}_{X,Y}[(Y-\mu(X))(\mu(X)-m(X))] &= \mathbb{E}_X[\mathbb{E}_{Y|X}(Y-\mu(X))(\mu(X)-m(X))\,|\,X] \\ &= \mathbb{E}_X[(\mu(X)-m(X))\,\underbrace{\mathbb{E}_{Y|X}[Y-\mu(X)|X]}_{=0 \text{ by defn of } \mu(X)}\end{aligned}$$

From $\mathbb{E}_{X,Y}[(Y-m(X))^2] = \mathbb{E}_{X,Y}[(Y-\mu(X))^2] + \mathbb{E}_X[(\mu(X)-m(X))^2]$,
it must be that MSE is minimized when $\mathbb{E}_X[(\mu(X)-m(X))^2]$ is minimized.

$\mathbb{E}((\mu(X)-m(X))^2] = 0$ when $m(X) = \mu(X)$:

$$\begin{aligned}m^*(X) &= \mu(X) \\ &= \mathbb{E}_{Y|X}(Y|X=x).\end{aligned}$$

Thus the expected MSE-optimal prediction is made by using $\mu(X)$.

We call $\mathbb{E}(Y|X=x)$ the regression function.

Lecture 3 (Sep 10) KNN Regression

How might we estimate the regression function?

$$\downarrow$$

$$m^*(x) = \mathbb{E}(Y \mid X = x)$$

Given n observations : (x_i, y_i) , $1 \le i \le n$, $x_1, \dots, x_n \in \mathbb{R}^p$

$\mathbb{E}(Y \mid X = x) = ?$

observations

- average? $\hat{m}(x) = \text{average}(\{y_i : x_i = x\})$
  ie for an x_i = some x value we set, take the mean of the y_i observations that correspond to $x_i = x$.

  → problem with this: what if there's only 1 point at $x_i = x$? Cannot estimate directly
  We can relax our criteria and use data points that are in the vicinity of the $x_i = x$ point.

- average of nearest neighbors:

$$\hat{m}(x) = \text{average}(\{y_i : x_i \text{ equal to or close to } x\})$$

$$= \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

- where $N_k(x)$ is the neighbourhood of x defined by the k-closest points x_i in the training sample.
  i.e. Euclidean distance.
- k is a hyper-parameter that we set

e.g. our dataset (n=6) , $X = (x_1, x_2, x_3)$, y

x_{i1}	x_{i2}	x_{i3}	y_i
1	1	0	3
2	0	3	2
0	1	2	0
0	1	3	1
-1	1	1	4
1	1	2	3

Idea: use KNN regression to make predictions about y.

⇒ what is $\hat{m}(x)$?

let $x_0 = (0, 0, 0)$, $y_0 = ?$

→ Compute the Euclidean dist. between each obs and x_0.

- $\|x_1 - x_0\|_2 = \sqrt{(1-0)^2 + (1-0)^2 + (0-0)^2}$
  $= \sqrt{2} \approx 1.41$
- $\|x_2 - x_0\|_2 = 3.61$
- $\|x_3 - x_0\|_2 = 2.23$
- $\|x_4 - x_0\|_2 = 3.16$
- $\|x_5 - x_0\|_2 = 1.73$
- $\|x_6 - x_0\|_2 = 2.44$

x_0 neighbours in order of closeness: $\{x_1, x_5, x_3, x_6, x_4, x_2\}$

our formula is $\hat{m}(x) = \frac{1}{k}\sum_{x_i \in N_k(x} y_i$

→ KNN prediction when $k=1$?

$N_1(x_0) = \{x_1\}$

$\hat{m}(x_0) = y_1 = \underline{3}$ //

→ KNN prediction when $k=3$?

$N_3(x_0) = \{x_1, x_5, x_3\}$

$$\hat{m}(x_0) = \frac{1}{3}\sum^{3}_{x_i \in N_3(x_0)} y_i$$

$$= \frac{1}{3}(y_1 + y_5 + y_3)$$

$$= \frac{1}{3}(3 + 0 + 4) = \underline{2.33}$$ //

Ⓡ R computation for KNN prediction in handouts

Lec 4 (Sep 15) Limitat$^{\underline{n}}$s of KNN

Theoretical guarantee of KNN:

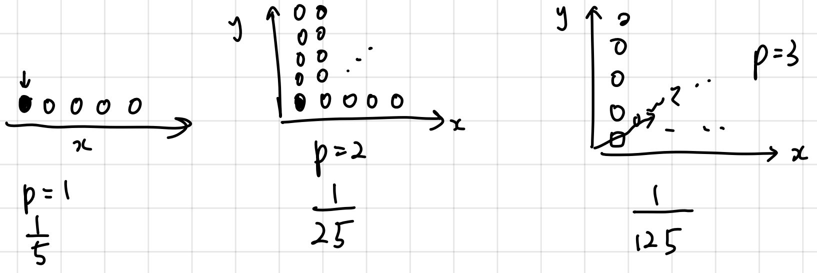
- $X \in \mathbb{R}^p$, $Y \in \mathbb{R}$
  ↖ vector of p variables, our predictors
- KNN can be pretty good for small p (i.e. $p<4$) and large N.

Under mild regularity conditions on joint prob. dist. of $P_r(X,Y)$, one can show that as $N, k \to \infty$ such that $\frac{k}{N} \to 0$ (i.e. k grows at a slower rate than N), then:

$$\hat{m}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i \longrightarrow \mathbb{E}(Y \mid X = x)$$

Curse of dimensionality

- as we increase the dimension of predictors, model performance is affected.
- for small p, we can always find a fairly large neighbourhood of observations close to the target x and take their average.
- But for very large p, it's more difficult to find a neighbourhood that contains enough data.

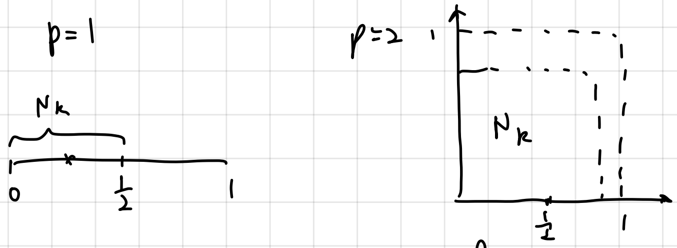


The "area" that a neighbourhood takes up in the data space grows smaller and smaller

- Consider the KNN for inputs X uniformly distributed in a p- dim unit hypercube

- Suppose we want to predict at a target X, set up a neighbourhood around X to capture a fraction R of the observation.

$R = \frac{1}{2}$



Expected edge length ("radius") of the neighborhood:

$$L_p(R) = R^{\frac{1}{p}}$$

eg $L_{10}(0.01) = 0.63$ → to capture 1% of data, in a 10-dimensional case, radius is 0.63

$L_{10}(0.1) = 0.8$ * In our eg, it's a unit, so 0.8 radius is almost the whole space. This is not local anymore!

↑ 10% of data

To capture 1% or 10% of the data to form a local average, we must cover 63% or 80% respectively of the range of each input variable. Such neighborhoods are not "local" anymore.

Optimal Linear Prediction

In general, $\hat{m}(x) = \mathbb{E}(Y \mid X = x)$ might be a very complicated form:

model	$\mathbb{E}(Y \mid X = x)$
KNN	$\frac{1}{k} \sum_{x_i \in N_k(x)} y_i$
linear regression	$x' \beta$
additive model	$f_1(x_1) + \cdots + f_p(x_p)$, $X = (x_1, \ldots, x_p)$
decision tree	$T(x)$ ← nonlinear function
random forest/ gradient boosting	$\sum_{m=1}^{M} \beta_m T_m(x)$
deep learning	$f(\sigma(W_x))$
SVM	$\sum_i \alpha_i K(X, X_i)$

To simplify m(x), we restrict m(x):

$$m(X) = \beta_0 + \beta_1 X$$

↖ only 1 predictor $X \in \mathbb{R}$

→ What are the optimal values of β_0 and β_1?

Theoretical ans:

- if distribution (X, Y) known

MSE

$$(\beta_0^*, \beta_1^*) = \underset{(\beta_0,\, \beta_1)}{\arg\min}\ \mathbb{E}_{X,Y}\left[(Y - m(X))^2\right]$$

$$= \underset{(\beta_0,\, \beta_1)}{\arg\min}\ \mathbb{E}_{X,Y}\left[Y - (\beta_0 + \beta_1 X))^2\right]$$

$$\Rightarrow \beta_1^* = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}, \quad \beta_0^* = \mathbb{E}(Y) - \beta_1^* \mathbb{E}(X)$$

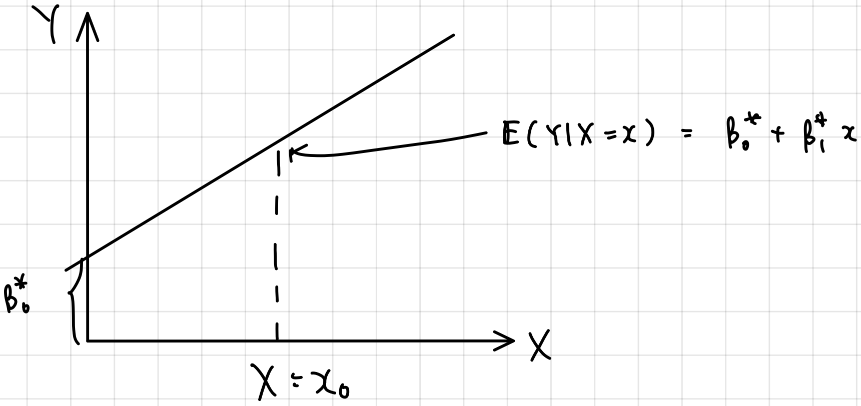
Proof is in handout:
1. decompose MSE
2. set derivs to 0
3. plug and solve

⇒ optimal linear predictor: $m^*(X) = \beta_0 + \beta_1 X$

$$= \mathbb{E}(Y) - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \mathbb{E}(X) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} X$$

Lec 5 (Sep 17)

The line $\beta_0^* + \beta_1^* X$ is called the optimal prediction line i.e. linear regression function.



Some notes about the optimal predictor line:

1. The optimal predutor line passes through $(\mathbb{E}(X), \mathbb{E}(Y))$

$$m^*(x) = \beta_0^* + \beta_1^* x$$

$$= \underbrace{\mathbb{E}(Y) - \beta_1^* \mathbb{E}(X)}_{\beta_0^*} + \beta_1^* x$$

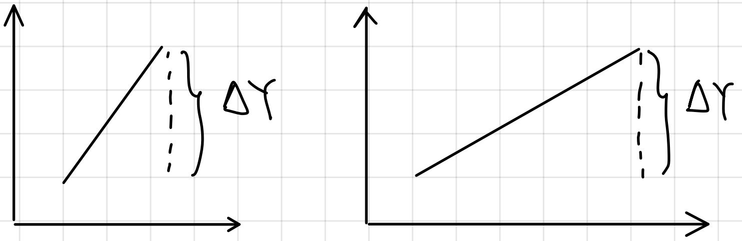
$$\Rightarrow m^*(\mathbb{E}(X)) = \mathbb{E}(Y)$$

2. If X and Y are "centered", i.e. $\mathbb{E}(Y) = \mathbb{E}(X) = 0$, the optimal regression line passes through $(0,0)$ since $\beta_0^* = 0$.

Recall that optimal slope $\beta_1^* = \dfrac{\text{Cov}(X,Y)}{\text{Var}(X)}$.

3. The optimal slope β_1^* increases as $\text{Cov}(X,Y)$ increases.

4. The optimal slope β_1^* decreases as $\text{Var}(X)$ increases.



For the same Δ in Y, larger variance of X leads to flatter slope.

5. The optimal slope β_1^* does not change if we use instead $Y - c$ and $X - c'$. But the intercept β_0^* will change.

6. Non-linear pattern cannot be appropriately modelled.

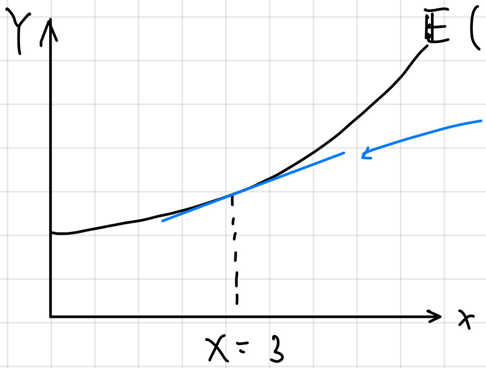
eg: Imagine true regression function:

$$\mathbb{E}(Y \mid X = x) = e^x \longleftarrow (\text{this is nonlinear})$$

Taylor's expansion at $X = x_0$

$$e^x = e^{x_0} + \left.\frac{\partial e^x}{\partial x}\right|_{x = x_0} (x - x_0) + \frac{1}{2} \frac{\partial^2 e^x}{\partial x^2} (x - x_0)^2 + \cdots$$

$$= e^{x_0} + e^{x_0}(x - x_0) + \frac{1}{2} e^{x_0} (x - x_0)^2 + \cdots$$



$\mathbb{E}(Y \mid X = x) = e^x$

$$x = 3 : \underbrace{e^3 + e^3(x-3)} + \frac{1}{2}e^3(x-3)^2 + \cdots$$

these first 2 terms are linear

- if x is close to 3, the linear prediction function is not bad!
- but if we are trying to predict say $x = 10$ then the function is not very good.

If we want to use the linear component as a prediction function, need to make sure that it dominates the quadratic term

i.e. $\frac{1}{2} e^{x_0} |x - x_0|^2 \ll e^{x_0} |x - x_0|$

$$\frac{|x-x_0|^2}{|x-x_0|} < \frac{2e^{x_0}}{e^{x_0}}$$

$$|x - x_0| < 2$$

$\Rightarrow$ In our example, $|x-3| < 2$

$$x - 3 < 2 \quad \text{or} \quad x - 3 > -2$$

$$x < 5 \quad \text{or} \quad x > 1$$

Linear prediction is good for $1 < x < 5$

Plug-in estimation (estimating optimal linear prediction using data)

How to estimate optimal linear prediction $m^*(x) = \mathbb{E}(Y \mid X = x) = \beta_0^* + \beta_1^* x$ from n observations of data $(x_1, y_1), \ldots (x_n, y_n)$?

$$\Rightarrow \hat{\beta}_1 = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\text{Var}}(X)}$$

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

$$\hat{\beta}_0 = \hat{\mathbb{E}}(Y) - \hat{\beta}_1 \hat{\mathbb{E}}(X)$$

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

The fitted regression line: $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

- an approximation of the true regression line, calculated with data.

$$\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$= \left[\bar{y} - \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - x)^2} \right] + \left(\frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) x$$

see handout for proof.

idea: plug-in sample values, i.e. $\left.\begin{array}{l} \mathbb{E}(Y) \approx \bar{y} \\ \mathbb{E}(X) \approx \bar{x} \end{array}\right\}$ sample mean

sample variance

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}X)^2) \approx \underline{\tfrac{1}{n}} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

we could use $\frac{1}{n-1}$ too for sample but it doesnt matter when n is large.

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y))$$

$$\approx \underline{\tfrac{1}{n}} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

sample covariance

some notation:

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

and $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i - \bar{x}) x_i = \sum_{i=1}^{n} x_i^2 - n(\bar{x})^2$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \frac{S_{xy}}{S_{xx}} \bar{x}$$

useful result: $\sum_{i=1}^{n} (y_i - \bar{y}) = 0$, $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

Some notes:

1. As $n \to \infty$, $\hat{\beta}_1 \longrightarrow \beta_1^*$ and $\hat{\beta}_0 \longrightarrow \beta_0^*$

2. The fitted regression line $\hat{m}(x)$ passes through $(\bar{x}, \bar{y})$.

3. If the data is centered, $x_i' = x_i - \bar{x}$, $y_i' = y_i - \bar{y}$ $\forall\, i = 1, \dots, n$

   $\bar{x}' = \frac{1}{n} \sum_{i=1}^{n} x_i' = 0$, $\bar{y}' = \frac{1}{n} \sum_{i=1}^{n} y_i' = 0$,

   then the fitted line passes through $(0, 0)$.

4. Slope does not change under a shift of the data

   ← eg $y_i' = y_i - c$, $x_i' = x_i - c'$

lec 6 (Sep 22) Simple Linear Regression

"simple": 1 predictor for 1 DV

We are interested in the underlying mechanism / DGP.

Recall 2 goals of regression:

1. prediction: forecast unobserved data (cause → effect)
2. inference: identify unknown data generating process

In order to do inference, we have to make assumptions and believe our data is generated from the SLR model.

→ we think that the model approximates reality relatively well.

SLR model specification:

- consider 2 rvs: Y (the response / DV), $X_1 \leftarrow$ only 1 predictor
- our goal here is to predict Y using X_1.
- assumptions:

A1. arbitrary input: The distribution of X_1 is arbitrary, and X_1 is even non-random.

A2. linear function and additive error: $Y = \underbrace{\beta_0 + \beta_1 X_1}_{\text{linear fcn}} + \underbrace{\varepsilon}_{\text{error}}$

e.g. if $X_1 = x_1$, then $Y = \beta_0 + \beta_1 x_1 + \varepsilon$, for unknown coefficients β_0 and β_1, and random noise ε.

A3. zero mean and constant variance error (homoskedasticity):

$E(\varepsilon) = 0$, $var(\varepsilon) = \sigma^2 \geq 0$., σ^2 is unknown

A4. independent error: ε is independent of X.

Some notes about SLR:

- if $\mathbb{E}(\varepsilon) = c \neq 0$, we can find an rv ε' with $\mathbb{E}(\varepsilon') = 0$ and $\text{var}(\varepsilon') = \sigma^2$, s.t. $\varepsilon = \varepsilon' + c$.

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \varepsilon \\
&= \beta_0 + \beta_1 X_1 + \varepsilon' + c \\
&= \underbrace{(\beta_0 + c)}_{\beta_0'} + \beta_1 X_1 + \varepsilon'
\end{aligned}
$$

- SLR assumptions actually imply that the true (optimal) regression line is linear and Y has constant variance.

$$
\begin{aligned}
\mathbb{E}(Y \mid X_1 = x_1) &= \mathbb{E}(\beta_0 + \beta_1 X_1 + \varepsilon \mid X_1 = x_1) \\
&= \mathbb{E}(\beta_0 + \beta_1 X_1 \mid X_1 = x_1) + \underbrace{\mathbb{E}(\varepsilon)}_{=0} \\
&= \beta_0 + \beta_1 x_1 \quad \leftarrow \text{this is a realized value}
\end{aligned}
$$

$$
\begin{aligned}
\text{Var}(Y \mid X_1 = x_1) &= \text{Var}(\beta_0 + \beta_1 X_1 + \varepsilon \mid X_1 = x_1) \\
&= \underbrace{\text{var}(\beta_0)}_{0} + \underbrace{\text{var}(\beta_1 x_1)}_{0} + \text{var}(\varepsilon) \\
&= \text{Var}(\varepsilon) \\
&= \sigma^2
\end{aligned}
$$

$$\Rightarrow Y \mid X_1 = x_1 \overset{d}{\sim} (\beta_0 + \beta_1 x_1 \, , \; \sigma^2)$$

Lec7 (Sep 24)

The noise variable ε can represent:

- other factors not considered in the model
- measurement error
- or some combination of both.

We think of the SLR assumptions A1 – A4 as modeling decisions that (we hope) will be useful, rather than as facts about the actual underlying relationship between Y and Xs.

Interpretation of parameters:

- β_0 = intercept, the expected value of Y when X_1 is 0.

$$\mathbb{E}(Y \mid X_1 = 0) = \beta_0 + 0 \cdot \beta_1 = \beta_0$$

- β_1 = slope, diff between exp. value of Y when x_1 is shifted by 1.

$$E(Y \mid X_1 = x_1) - E(Y \mid X_1 = x_1 + 1) = \beta_0 + \beta_1 x_1 - \beta_0 - \beta_1 x_1 - \beta_1$$
$$= \beta_1$$

Note that β_1 does not imply causality! It is only statistical association.

If we select 2 sets of cases from (X_1, Y) distribution where X_1 differs by 1, we expect the associated Y to differ by β_1.

- σ^2 = error variance, the variance of the noise around the reg. line. It represents a typical distance of a point from the true regression line.

Model set up for multiple data points:

We assume multiple data points $(X_{11}, Y_1), (X_{21}, Y_2), \dots, (X_{n1}, Y_n)$ are generated from the same model.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i, \quad i = 1, \dots, n$$

where $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2 \quad \forall i$,

and ε_i and ε_j are independent for $i \neq j$

Equivalently, $E(Y_i \mid X_{i1} = x_{i1}) = \beta_0 + \beta_1 x_{1i}$

$$Var(Y_i \mid X_{i1} = x_{i1}) = \sigma^2, \quad i = 1, \dots, n$$

Question: $Y_i \perp Y_j$ (independent)?

$Y_i \perp Y_j \mid X_{i1}, X_{i2}$ (conditionally independent)?

Y_i and Y_j are not unconditionally independent:

$$Cov(Y_i, Y_j) = Cov(\beta_0 + \beta_1 X_i + \varepsilon_i, \ \beta_0 + \beta_1 X_j + \varepsilon_j)$$
$$= \beta_1^2 \, cov(X_i, X_j) + \beta_1 \underbrace{cov(X_i, \varepsilon_j)}_{=0} + \beta_1 \underbrace{cov(\varepsilon_i, x_j)}_{=0} + \underbrace{cov(\varepsilon_i, \varepsilon_j)}_{=0}$$

because X and ε are indep.

$$= \beta_1^2 \, cov(X_i, X_j)$$

If $X_i \perp X_j$, $cov(X_i, X_j) = 0 \Rightarrow Cov(Y_i, Y_j) = 0$

But if not $X_i \perp X_j$, $cov(Y_i, Y_j) \neq 0$

Y_i and Y_j are conditionally independent given X_i, X_j:

$$\begin{aligned} \mathrm{Cov}(Y_i, Y_j \mid X_i = x_i, X_j = x_j) &= \mathrm{Cov}(\varepsilon_i, \varepsilon_j \mid X_i = x_i, X_j = x_j) \\ &= \mathrm{Cov}(\varepsilon_i, \varepsilon_j) \\ &= 0 \end{aligned}$$

← as realized values, they become constants and $\mathrm{cov} = 0$.

Ⓡ Generate data from SLR model (simulation)

Optimal prediction for SLR

Assume (X_1, Y) are generated from the SLR model:

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon, \quad \mathbb{E}(\varepsilon) = 0, \quad \mathrm{Var}(\varepsilon) = \sigma^2, \quad \varepsilon \perp X_1$$

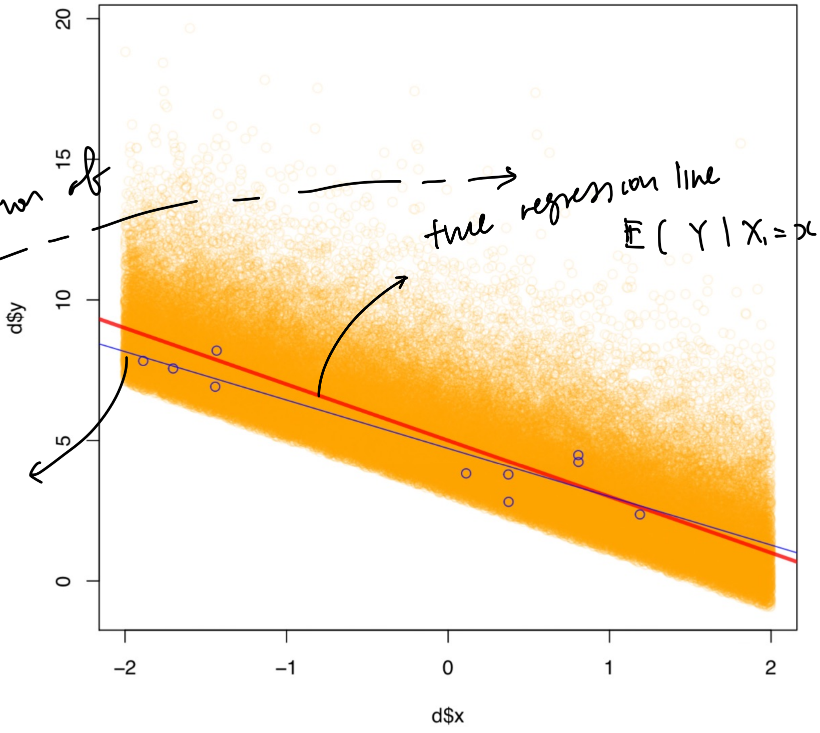
If we want to predict Y using X_1, what is the optimal prediction that minimizes the mean squared error?

$$m^*(\cdot) = \underset{m(\cdot)}{\mathrm{argmin}} \ \mathbb{E}_{X,Y}[(Y - m(X_1))^2]$$

$\Rightarrow m^*(x_1) = \mathbb{E}(Y \mid X_1 = x_1)$ is the optimal prediction function

$$\begin{aligned} &= \mathbb{E}(\beta_0 + \beta_1 X_1 + \varepsilon \mid X_1 = x_1) \\ &= \underbrace{\beta_0 + \beta_1 x_1}_{\text{true regression line}} \end{aligned}$$

Lec 8 (Sep 29)



an approximation of

true regression line

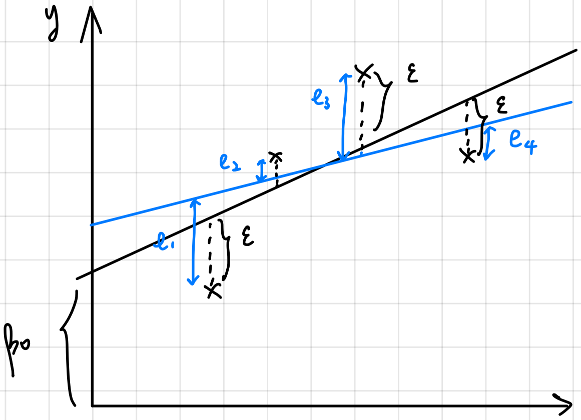
$\mathbb{E}(Y \mid X_1 = x_1) = \beta_0 + \beta_1 x_1$

fitted regression line

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

Least Squares estimators

Given n observations (x_{11}, y_1), (x_{21}, y_2), $\dots$, (x_{n1}, y_n) to estimate β_0 and β_1 in SLR.



$\mathbb{E}(Y \mid X_1 = x_1) = \beta_0 + \beta_1 x_1$

Goal: minimize the residual sum of squares

Residual sum of squares: $\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta} x_{i1})^2$

Least squares solve $\hat{\beta}_0$, $\hat{\beta}_1$ such that

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} e_i^2$$
$$= \underset{\beta_0, \beta_1}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})^2$$

We can show that:

$$\left.\begin{aligned} \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_{i1} - \bar{x})}{\sum_{i=1}^{n}(x_{i1} - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}_1 \end{aligned}\right\}$$

the least square estimators are the same as the plug in estimators

Ideally, $(\beta_0^*, \beta_1^*) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \underbrace{E_{X,Y}\left[(Y - \beta_0 - \beta_1 X_1)^2\right]}_{\text{"expected risk"}}$

the least square estimators are a sample estimation of the above:

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1})^2}_{\text{"empirical risk"}}$$

→ Proof: $S(\beta_0, \beta_1) = \widehat{MSE}(\beta_0, \beta_1)$

$$= \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1})^2$$

$$\frac{\partial S}{\partial \beta_0} = \frac{1}{n}(-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1}) := 0$$

$$\frac{\partial S}{\partial \beta_1} = \frac{1}{n}(-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1})\, x_{i1} := 0$$

$$\Rightarrow \left\{\begin{aligned} &\bar{y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{x}_1 = 0 \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 \\ &\frac{1}{n}\sum_{i=1}^{n} x_{i1} y_i - \hat{\beta}_0 \bar{x}_1 - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_{i1}^2 = 0 \end{aligned}\right.$$

Substitute $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1$ to obtain $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$. ▨

Ⓡ See handout for least squares in R.

SLR in matrix form

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \underset{\sim}{X} = \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{n1} \end{bmatrix}, \quad \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

↑ not observed yet, all are R.V.s

Assumptions (recall SLR A1 - A4):

A1. same — arbitrary input

A2. $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

A3. $\mathbb{E}(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$, $\text{var}(\underset{\sim}{\varepsilon}) = \sigma^2 \underset{\sim}{I}_n$

A4. implied by A3 — independence of error ($\underset{\sim}{\varepsilon} \perp \underset{\sim}{X}$)

$$\begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}$$

variance - covariance matrix:

$$\begin{bmatrix} \text{cov}(\varepsilon_1, \varepsilon_1) & \text{cov}(\varepsilon_1, \varepsilon_2) & \cdots \\ \text{cov}(\varepsilon_2, \varepsilon_1) & \text{cov}(\varepsilon_2, \varepsilon_2) & \cdots \\ \vdots & \ddots & \end{bmatrix}$$

- $\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$ as $\varepsilon_i \perp \varepsilon_j$
- $\text{cov}(\varepsilon_i, \varepsilon_i) = \text{var}(\varepsilon_i) = \sigma^2$

Some notes:

1. $\mathbb{E}(\underset{\sim}{Y} \mid \underset{\sim}{X}) = \underset{\sim}{X}\underset{\sim}{\beta} + \mathbb{E}(\underset{\sim}{\varepsilon})$

$$= \underset{\sim}{X}\underset{\sim}{\beta}$$

In regular form, $\mathbb{E}(Y_1 \mid X_1 = x_1) = \beta_0 + \beta_1 x_1$ for $i = 1, \dots, n$

2. $\underset{\sim}{Y}$ has constant variance: $\text{Var}(\underset{\sim}{Y} \mid \underset{\sim}{X}) = \sigma^2 I_n = \text{var}(\underset{\sim}{\varepsilon})$

$$\text{var}(\underset{\sim}{Y} \mid \underset{\sim}{X}) = \begin{bmatrix} \text{var}(Y_1 \mid \underset{\sim}{X}) & \text{cov}(Y_1, Y_2 \mid \underset{\sim}{X}) & \cdots \\ \text{cov}(Y_2, Y_1 \mid \underset{\sim}{X}) & \text{var}(Y_2 \mid \underset{\sim}{X}) & \\ \vdots & & \ddots \end{bmatrix}$$

$\Rightarrow \text{var}(Y_i \mid X_{i1} = x_{i1}) = \sigma^2$ for $i = 1, \dots, n$

$\text{cov}(Y_j, Y_k \mid X_{j1} = x_{j1}, X_{k1} = x_{k1}) = 0$ for $j \neq k$

Least squares in matrix form

$$\hat{\underset{\sim}{\beta}} = \underset{\underset{\sim}{\beta}}{\text{argmin}} \ \frac{1}{n} \| \underset{\sim}{Y} - \underset{\sim}{X}\underset{\sim}{\beta} \|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2$$

→ solution is $\hat{\beta} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1} \underset{\sim}{X}'\underset{\sim}{Y}$ (proof in handout)

Lec 9 (Oct 1)

$$(\beta_0^*, \beta_1^*) = \underset{\beta_0, \beta_1}{\text{argmin}} \underbrace{E_{X,Y}\left[(Y - \beta_0 - \beta_1 X_1)^2\right]}_{\text{"expected risk"} = MSE(\beta_0, \beta_1)}$$

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\text{argmin}} \underbrace{\frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1})^2}_{\text{"empirical risk"} = \hat{MSE}(\beta_0, \beta_1)}$$

Note: If the data (x_{11}, y_1), (x_{21}, y_2), ... (x_{n1}, y_n) are all independent for any fixed, the Law of Large Numbers tells us that:

as $n \to \infty$, $\hat{MSE}(\beta_0 \cdot \beta_1) \longrightarrow MSE(\beta_0, \beta_1)$

If the SLR assumptions are true, then:

as $n \to \infty$, $(\hat{\beta}_0, \hat{\beta}_1) \longrightarrow (\beta_0^*, \beta_1^*)$

(i.e. approximation is good when n large)

For SLR, $Y = \beta_0 + \beta_1 X + \varepsilon$, optimal (true) regression function is:

$$E(Y \mid X = x) = \beta_0 + \beta_1 X$$
$$= \beta_0^* + \beta_1^* X$$

From these results, we see that not only β_0, β_1 are the parameters of the SLR, but also β_0, β_1 can minimize the expected risk (MSE), i.e. $\beta_0^* = \beta_0$, $\beta_1^* = \beta_1$

These symptotic results depend on SLR assumptions being true.

Statistical properties of least Squares

Estimates vs estimators?

- An estimate is a number (deterministic), the realized value of the estimator.
- An estimator is an r.v. (stochastic)

- For example, we can potentially draw n samples $X_1, X_2, \ldots, X_n$ from a distribution. The X_i's have not actually been observed yet and can potentially be any value from the distribution.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$ is also an r.v., whose distribution is a sampling distribution.

↑ this is an estimator of the population mean μ.

If the values are realized, then $\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$ is an estimate.

Least Square estimators

Given that the SLR model assumptions A1 – A4 are satisfied, we can show that the least square estimators are unbiased:

$$\mathbb{E}(\hat\beta_1) = \beta_1 , \quad \mathbb{E}(\hat\beta_0) = \beta_0$$

and variance of the estimators are:

$$Var(\hat\beta_1 | x_{11}, \dots, x_{n1}) = \frac{\sigma^2}{S_{xx}} , \quad var(\hat\beta_0 | x_{11}, \dots, x_{n1}) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_1^2}{S_{xx}} \right)$$

related to var of ε

proof in handout

Bias and variance in matrix form:

$$\underset{\sim}{Y} = \underset{\sim}{X} \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

$$\mathbb{E}(\underset{\sim}{Y} | \underset{\sim}{X}) = \underset{\sim}{X} \underset{\sim}{\beta}$$

$$Var(\underset{\sim}{Y} | \underset{\sim}{X}) = \sigma^2 \underset{\sim}{I}_n$$

unbiasedness: $\mathbb{E}(\hat{\underset{\sim}{\beta}} | \underset{\sim}{X}) = \underset{\sim}{\beta}$

variance : $var(\hat{\underset{\sim}{\beta}} | \underset{\sim}{X}) = \sigma^2 (\underset{\sim}{X}' \underset{\sim}{X})^{-1}$

$$\begin{bmatrix} Cov(\hat\beta_0, \hat\beta_0) & cov(\hat\beta_0, \hat\beta_1) \\ cov(\hat\beta_1, \hat\beta_0) & cov(\hat\beta_1, \hat\beta_1) \end{bmatrix} = \begin{bmatrix} var(\hat\beta_0) & cov(\hat\beta_0, \hat\beta_1) \\ cov(\hat\beta_1, \hat\beta_0) & var(\hat\beta_1) \end{bmatrix}$$

Lec 10 (Oct 6)

How to estimate σ^2:

Recall $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, $\mathbb{E}(\varepsilon) = 0$, $Var(\varepsilon) = \sigma^2$

Although σ^2 is not used to estimate $\hat\beta_0$ and $\hat\beta_1$, it is still important as it tells us about:

1. randomness of Y
2. σ^2 is related to $var(\hat\beta_0 | \mathbb{X})$ and $var(\hat\beta_1 | \mathbb{X})$

$x_{11}, x_{21}, \dots, x_{n1}$

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

$$Var(\hat\beta_1 | X) = \frac{\sigma^2}{S_{xx}}$$

$$\hat\sigma_2 = \frac{\text{sum of squared residuals}}{n-2} = \text{mean sq. residuals}$$

sum of sq. residuals: $SSR = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$

$= \sum_{i=1}^n (y_i - \hat\beta_0 - \hat\beta_1 x_{i1})^2$

mean sq. res: $MSR = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat\beta_0 - \hat\beta_1 x_{i1})^2 = \frac{1}{n-2} SSR$

recall that $\widehat{MSE} = \frac{1}{n} SSR$.

When n is large, $MSR \approx \widehat{MSE}$.

Recall that in SLR:

$$\begin{aligned}\mathbb{E}((Y-\beta_0-\beta_1X_1)^2) &= \mathbb{E}(\varepsilon^2)\\ &= Var(\varepsilon) + (\mathbb{E}(\varepsilon))^2\\ &= \sigma^2\end{aligned}$$

Plug-in principle: use the observed sample values

- Replace X_1 with $(x_{11}, x_{21}, \cdots, x_{n1})$
- Replace Y with $(y_1, y_2, \cdots, y_n)$
- $\mathbb{E}(\cdot)$ replaced by $\frac{1}{n}\sum_{i=1}^{n}$

$$\Rightarrow \mathbb{E}((Y-\beta_0-\beta_1X_1)^2) \approx \frac{1}{n}\sum_{i=1}^{n}[(y_i-\hat{\beta}_0-\hat{\beta}_1x_{1i})^2]$$ ← which is fully computable

$$\approx \frac{1}{n-2}\sum_{i=1}^{n}[(y_i-\hat{\beta}_0-\hat{\beta}_1x_{1i})^2]$$

↖ this adjustment makes it unbiased. Does not really matter when n is large, but makes a diff when n small.

$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ = residual standard error, standard error of regression

$\hat{\sigma}^2$ is unbiased: $\mathbb{E}(\hat{\sigma}^2) = \mathbb{E}\left(\frac{SSR}{n-2}\right) = \sigma^2$ (proof in notes)

↖ an r.v.

If we use n instead of $n-2$, i.e. $\tilde{\sigma}^2 = \frac{1}{n}(SSR)$, $\mathbb{E}(\tilde{\sigma}^2) = \frac{n-2}{n}\sigma^2 \xrightarrow{n\to\infty} \sigma^2$

Alternative formula for SSR: $SSR = \underbrace{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}_{TSS} - \hat{\beta}_1 S_{xy}$

total sum of squares: $TSS = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$

Using $\hat{\sigma}^2$ to estimate $Var(\hat{\beta}_0|\mathbb{X})$ and $Var(\hat{\beta}_1|\mathbb{X})$

$$\widehat{Var}(\hat{\beta}_1|\mathbb{X}) = \frac{\hat{\sigma}^2}{S_{xx}} \qquad \widehat{Var}(\hat{\beta}_0|\mathbb{X}) = \hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}_1^2}{S_{xx}}\right)$$

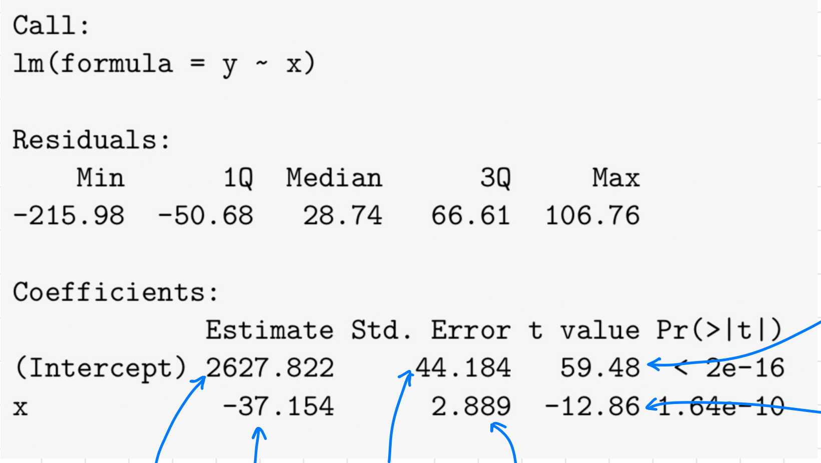
The standard errors (se) are the square roots of the variances.

$$se(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{S_{xx}}} \qquad se(\hat{\beta}_0) = \sqrt{\sigma^2\left(\frac{1}{n}+\frac{\bar{x}_1^2}{S_{xx}}\right)}$$

Replace σ^2 with $\hat{\sigma}^2$ for the estimates:

$$\widehat{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \qquad \widehat{se}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}_1^2}{S_{xx}}\right)}$$

Ⓡ estimation of σ^2



```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-215.98  -50.68   28.74   66.61  106.76

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2627.822     44.184   59.48  < 2e-16
x             -37.154      2.889  -12.86 1.64e-10
```

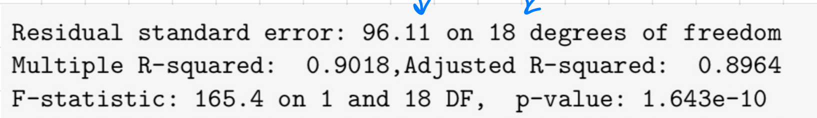
$T_0 = \frac{\hat{\beta}_0}{\hat{se}(\hat{\beta}_0)}$

$T_1 = \frac{\hat{\beta}_1}{\hat{se}(\hat{\beta}_1)}$

β_0 β_1 $\hat{se}(\hat{\beta}_0)$ $\hat{se}(\hat{\beta}_1)$

$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$

$n-2=18 \Rightarrow n=20$



```
Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared:  0.9018,Adjusted R-squared:  0.8964
F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10
```

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

Lec 11 (Oct 8)

+ 1 assumption : Gaussian-Noise Simple Linear Regression

We can show that :

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

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

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

$(p=2)$ ← 1 predictor 1 intercept

GN-SLR assumptions:

A1. ____________

2. ____________

3. $\varepsilon \sim N(0, \sigma^2)$

4. ____________

← same as SLR assumptions.

Note the additional Gaussian assumption on the distribution of ε (A3).

	plug in	LS	prediction	unbiased estimators	$\hat{\beta} \sim N$	t test F test	CI	R^2
SLR	✓	✓	✓	✓				✓
GN-SLR	✓	✓	✓	✓	✓	✓	✓	✓

The GN-SLR model is strictly stronger than SLR. This means that everything we have done so far directly applies to GN-SLR.

Sampling distribution of $\frac{\hat{\beta}_1-\beta_1}{se(\hat{\beta}_1)}$ and $\frac{\hat{\beta}_1-\beta_1}{\widehat{se}(\hat{\beta}_1)}$

Results: $\frac{\hat{\beta}_1-\beta_1}{se(\hat{\beta}_1)} \sim N(0,1)$ $\qquad \frac{\hat{\beta}_1-\beta_1}{\widehat{se}(\hat{\beta}_1)} \sim t_{n-p}$ ← # of variables

for $Y = \beta_0 + \beta_1 X_1$, $p=2$

$\Rightarrow \sim t_{n-2}$

Explanation:

Recall results from prev lecture: $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$

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

For an r.v. Z if $Z \sim N(\mu, \sigma^2)$, then $\frac{Z-\mu}{\sigma} \sim N(0,1)$. We can similarly standardize $\hat{\beta}_1$:

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

↑ this is not an observed value, it is unknown as it is related to the unknown σ^2

Replace it with estimated se: $\widehat{se}(\hat{\beta}_1)$

$$T_1 = \frac{\hat{\beta}_1-\beta_1}{\widehat{se}(\hat{\beta}_1)} \sim t_{n-2} \qquad T_0 = \frac{\hat{\beta}_0-\beta_0}{\widehat{se}(\hat{\beta}_1)} \sim t_{n-2}$$

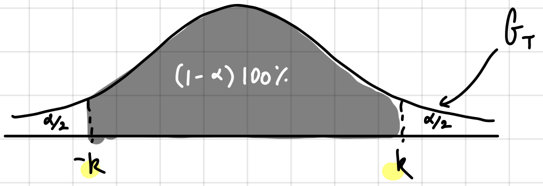
where $\widehat{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$ $\qquad \widehat{se}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2(\frac{1}{n}+\frac{\bar{x}_1^2}{S_{xx}})}$

$= \sqrt{\frac{MSR}{S_{xx}}}$ $\qquad = \sqrt{MSR(\frac{1}{n}+\frac{\bar{x}_1^2}{S_{xx}})}$

CI for β_1, β_0

Suppose that f_T is the density function of $T \sim t_{n-2}$.

Let $k>0$ s.t. $P(-k<T<k) = 1-\alpha$, $\alpha \in (0,1)$:



In the textbook,

$k \equiv t_{\frac{\alpha}{2}, \underbrace{n-2}_{\text{degrees of freedom}}}$

Result:

A $100(1-\alpha)\%$ confidence interval for β_1 is: ← not $\hat{\beta}_1$!!

$$CI(\beta_1) = [\hat{\beta}_1 - k\cdot\widehat{se}(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + k\cdot\widehat{se}(\hat{\beta}_1)]$$

→ Proof: $P(\beta_1 \in CI(\beta_1)) = P(\hat{\beta}_1 - k\cdot\widehat{se}(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + k\cdot\widehat{se}(\hat{\beta}_1))$

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

$$= P(-k \le T_1 \le k)$$

$$= 1-\alpha \quad \text{(by defn)}$$

The upper and lower bounds of $CI(\beta_1)$ are random as $\hat{\beta}_1$ can change if a different sample is drawn.
However, β_1 is fixed / non-random. ← unless you're a Bayesian

Explanation: The random interval $CI(\beta_1)$ traps β_1 with probability $1-\alpha$.
(i.e. If $\alpha = 0.05$, then 95% of the intervals of $CI(\beta_1)$, obtained by repeatedly sampling the data and finding $\hat{\beta}_1$, will contain β_1)

How does the width of $CI(\beta_1)$ change? width $= 2k \cdot \hat{se}(\hat{\beta}_1)$

- $\alpha \downarrow \Rightarrow (1-\alpha) \uparrow \Rightarrow$ width $\uparrow$
  High confidence comes at a price of big margin of error.
- $n \uparrow \Rightarrow$ width $\downarrow$
  Larger sample gives more accurate estimation.
- $\sigma^2 \uparrow \Rightarrow \hat{se}(\hat{\beta}_1) \uparrow \Rightarrow$ width $\uparrow$
  The more noise here is around the true regression line, the less precisely we can measure this line from the data.
- S_{xx} (the variation of X) $\uparrow \Rightarrow$ width $\downarrow$

Ⓡ simulation for constructing CI

Hypothesis Testing

Lec 12 · Oct 13

t Test (Wald Test) for β_1 using sampling distribution of β_1:

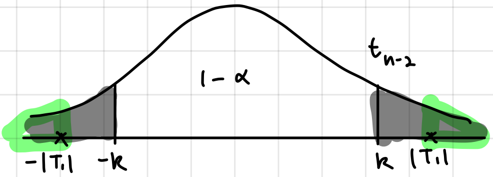
Suppose we want to test (2 sided test):

$H_0: \beta_1 = c$
$H_1: \beta_1 \neq c$

where c is a specific value e.g $c = 0$

method 1: Suppose we want to test with a significance level α, we compute the test statistics: $T_1 = \dfrac{\hat{\beta}_1 - c}{\hat{se}(\hat{\beta}_1)}$ ← all these values are known / can be computed

Reject H_0: $|T_1| \geq k \equiv t_{\frac{\alpha}{2}, n-2}$
$k(n, a)$ is $(1-\frac{\alpha}{2})$ 100% percentile of the distribution



or equivalently, consider the tail probability:

$$P(|T| > |T_1|) = \underbrace{P(T < -|T_1| \text{ or } T > |T_1|)}_{\text{p-value}} < \alpha$$

Oct 15

Why it works:

- We assume GN-SLR assumptions are satisfied.

If we assume that $H_0: \beta_1 = c$ is true, then we know the sampling dist for $T_1 = \frac{\hat{\beta}_1 - \beta_1}{\hat{se}(\hat{\beta}_1)}$ will be equal to $T_1 = \frac{\hat{\beta}_1 - c}{\hat{se}(\hat{\beta}_1)}$ and $T_1 \sim t_{n-2}$.

If we observe the unlikely result $|T_1| \geq k$, then there might be 2 explanations:

1. either in fact $\beta_1 = c$ (H_0 is true) and therefore you just by chance observe a very rare event that $|T_1| \geq k$ with the probability being only α
2. or the assumption is wrong and $\beta_1 \neq c$ (H_0 is not true) and H_0 should be rejected.

Method 2: We will reject H_0 in a hypothesis test with significant level α if the $100(1-\alpha)\%$ confidence interval $CI(\beta_1)$ does not cover c, and do not reject H_0 if $CI(\beta_1)$ contains c.

Test of significance of regression

$$H_0: \beta_1 = 0 = c$$
$$H_1: \beta_1 \neq 0$$

If we fail to reject H_0, this implies that there is no linear relationship between X and Y.

Comments on hypothesis testing:

- what α should we use? conventionally $\alpha = 0.05$
  ↖ false discovery rate

  If you are very conservative and cant afford false rejection, should use smaller α.
- statistical significance:

  If we test $H_0: \beta_1 = 0$ and we reject it, then we say the difference between β_1 and 0 is statistically significant with a given significance level α.

Permutation test ($H_0: \beta_1 = 0$; $H_1: \beta_1 \neq 0$)

↖ can only test for $c = 0$ in permutation test!

Steps:

1. compute the observed value of the t statistic: $T_1 = \frac{\hat{\beta}}{\hat{se}(\beta_1)}$
2. randomly permute the data $Y_1, \dots, Y_n$ while keeping $X_1 \dots X_n$ unchanged. Recompute the statistic T_1 again using permuted data.
3. repeat the previous step B times and let $Z_1 \dots Z_B$
4. The approximated p-value is p-value $= \frac{1}{B} \sum_{j=1}^{n} \mathbb{I}(|Z_j| > |T_1|)$

   reject $H_0: \beta_1 = 0$ if p-value $< \alpha$

   ↖ indicator function

- small permuted p-value is a strong indication of the rejection of the null hypothesis.
- permuted p-value also indicates false discovery rate
- idea of permutation test is trying to see if there is an association between Y and X. That's why we "scramble" the Ys.

Analysis of variance · Oct 20

SS_T (total sum of squares) : measures the total variation in Y.

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

SS_R (regression sum of squares) : measures amount of "systematic variation" in Y due to the Y ~ X linear relationship.

↳ i.e. the variation in Y that can be explained by the regression model.

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

↑ pred. value of y

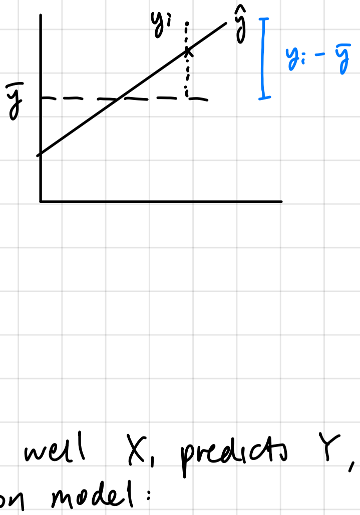
SS_{Res} (residual sum of squares) : measures amount of "residual variation" in Y. Aggregate measure of misfit of the regression line

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

In general, $SS_T = SS_R + SS_{Res}$

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

→ Proof in handout.



For observation i : $y_i - \bar{y} = \hat{y}_i - y_i + y_i - \bar{y}$

When we have a perfect fit, $\hat{y}_i = y_i$

$$SS_{Res} = 0 \Rightarrow SS_T = SS_R$$

R^2 : goodness of fit · Oct 22

If we want to define a global measure of how well X_1 predicts Y, we consider the proportion of variation explained by the regression model:

$$R^2 = \frac{SS_R}{SS_T} = \frac{SS_R}{SS_R + SS_{Res}} = 1 - \frac{SS_{Res}}{SS_T}$$

We can show that $SS_R = \hat{\beta}_1 S_{xy}$ (see handout)

$R^2 \in [0, 1]$.

- $R^2 = 0$ when $\hat{\beta}_1 = 0$
  the model cannot explain any variation in Y, very bad fit
- $R^2 = 1$ when $SS_{Res} = 0$
- If R^2 is near 1 then the predictor X_1 can explain a large proportion of the observed variation in Y
  → i.e. the predictor X_1 included in the model is a "sufficient predictor" of Y.

R^2 vs t-test on β_1? $(H_0: \beta_1 = 0 \,;\, H_1: \beta_1 \neq 0)$

- R^2 tells us how much variation is explained by including X_1 (is X_1 a sufficient pred. of Y?)
- t-test tells us about whether X_1 is necessary in explaining the variation of Y.

We consider 4 scenarios:

1. insignificant p-value and low R^2
   - X_1 is not useful AND the model does not explain much of the variation (worst case)
   - $\Rightarrow \beta_1 = 0$
2. insignificant p-value and high R^2
   - X_1 not useful and the model explains a lot of variation within the data.
     i.e. model without X_1 is already sufficient
3. significant p-value and low R^2
   - X_1 is at least useful but not sufficient, should add more predictors.
4. significant p-value and high R^2
   - X_1 is useful and sufficient (best case)

Adjusted R^2

# of variables in model

e.g in SLR, $p=2$ (β_0, β_1)

We have $E_{Y|X}(SS_{Res} \mid \mathbb{X}) = \sigma^2 (n-p)$

If we increases the model "complexity" to n, i.e. $p \to n$, then

$$SS_{Res} \xrightarrow{p \to n} 0$$

$$\Rightarrow R^2 = 1 - \frac{SS_{Res}}{SS_T} \xrightarrow{p \to n} 1$$

Overfitting!

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

- $\uparrow p \Rightarrow \uparrow$ numerator $\Rightarrow \downarrow R^2_{adj}$

R^2_{adj} can account for model complexity.

Relationship between R^2 and R^2_{adj}: $R^2 = 1 - \dfrac{SS_{Res}}{SS_T} = 1 - \dfrac{n-p}{n-1}(1 - R^2_{adj})$

- intercept only model $(p=1)$: $R^2_{adj} = R^2$
- SLR model $p=2$: $R^2_{adj} \neq R^2$
- when we fix p, $n \to \infty$, $R^2_{adj} \xrightarrow{n \to \infty} R^2$
  If $n >> p$, safe to use R^2. There is no overfitting issue when n is large.

Limitations of R^2

As $n \to \infty$, what does R^2 converge to?

$$R^2 = \frac{SS_R}{SS_T} = \frac{SS_R}{SS_R + SS_{Res}} = \frac{\hat{\beta}_1^2 S_{xx}}{\hat{\beta}_1^2 S_{xx} + SS_{Res}}$$

$$= \frac{\hat{\beta}_1^2 S_{xx} / n}{\hat{\beta}_1^2 S_{xx}/n + SS_{Res}/n}$$

As $n \to \infty$, $\hat{\beta}_1 \to \beta_1$, $\frac{S_{xx}}{n} \to Var(X_1)$, $\frac{SS_{Res}}{n} \to \sigma^2$

Thus, $R^2 \xrightarrow{n\to\infty} \frac{\beta_1^2 \, var(X_1)}{\beta_1^2 \, var(X_1) + \sigma^2}$ if SLR assumptions hold

- By making $var(X_1)$ small or σ^2 large, we can drive R^2 towards 0 even if the model is correct.
- Conversely, even if the model is wrong, we can make R^2 close to 1 by increasing $var(X_1)$ or making σ^2 small
- R^2 can be compared only when different models are fit to the same dataset. We cannot compare R^2 across diff datasets!
- R^2 is not as useful as a goodness of fit measure on training data, but more useful for testing data.

F test and ANOVA · Oct 27

The idea is to compare 2 models: (what does data look like?)

$$\left.\begin{array}{ll} H_0: Y = \beta_0 + \varepsilon & \\ H_1: Y = \beta_0 + \beta_1 X_1 + \varepsilon & (SLR) \end{array}\right\} \quad \text{equivalently. (t test on } \beta_1) \quad \begin{array}{l} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{array}$$

Constructing the F test:

idea: If we fit the first model, the OLS estimator is $\hat{\beta}_0 = \bar{y}$.
The idea of the F test is to create a test statistic that measures how much better the 2rd model is compared to the 1st model.

Under GN-SLR assumption, and if H_0 is true (i.e. $\beta_1 = 0$), we can show that

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

$$\frac{SS_{Res}}{\sigma^2} \sim \chi^2_{n-p}.$$ In this case, $p=2$, so $df_{Res} = n-2$

we refer to the model in the alt. hyp. EVEN THOUGH we assume Ho true!

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

For SLR, $p=2$. We will generalize this later.

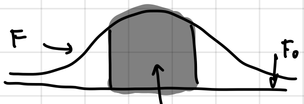
Degrees of freedom are additive:
$$\begin{array}{ccccc} df_T & = & df_R & + & df_{Res} \\ n-1 & = & p-1 & + & n-p \end{array}$$

Consider the F-test statistic:

$$F_0 = \frac{SS_R / df_R}{SS_{Res} / df_{Res}} = \frac{SS_R / p-1}{SS_{Res} / n-p} \equiv \frac{\overset{\text{mean sq.}}{MS_R}}{MS_{Res}}$$

Under $H_0 : \beta_1 = 0$, $F_0 \sim F_{(p-1, n-p)} = F_{(1, n-2)}$ for SLR

Note that $F_0 \geq 0$. So we only need to do a 1-sided test.
We reject $H_0 : \beta_1 = 0$ if F_0 is large enough

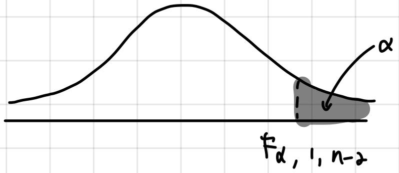


F_0 extreme for its dist! ⇒ H_0 might be wrong.

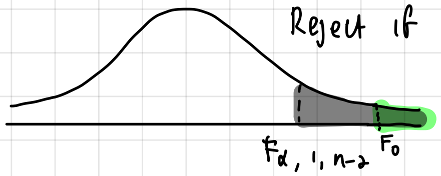
we expect F_0 to be here if H_0 true.

Set $\alpha = 0.05$, 0.01 etc. Reject H_0 if $\overset{\text{observed statistic}}{F_0} > F_{\alpha, p-1, n-p}$.
$F_{\alpha, 1, n-2}$ represents the $1-\alpha$ quantile of the F distribution with df 1 and n-2.

$$\text{i.e. } P(F > F_{\alpha, 1, n-2}) = \alpha$$



Equivalently, we can do F-test using p-value. p-value = $P(F > F_0)$



Reject if p-value < α

We summarize the F-test in an ANOVA table:

	SS	df	MS	F
Regression	SS_R	$p-1$	$MS_R = \frac{SS_R}{p-1}$	$F_0 = \frac{MS_R}{MS_{Res}}$
Residual	SS_{Res}	$n-p$	$MS_{Res} = \frac{SS_{Res}}{n-p}$	
Total	SS_T	$n-1$		

R use anova() function

Some notes on F-test

1. In deriving the F distribution, it is absolutely vital that all GN-SLR assumptions hold. The test never doubts that the right model is linear.

2. If we dont reject Ho, we dont find any significant share of variance associated with the regression.

$$F_0 = \frac{SS_R / df_R}{SS_{Res} / df_{Res}} \Big\} \text{ larger}$$

The following may be interpreted:

a) the intercept only model is better

b) $\beta_1 \neq 0$ but the data doesnt provide enough power to detect departure from the null.
↳ this is a rather conservative expl^n.

c) the real relationship between X_1 and Y is nonlinear but the best approximation to it has slope zero.
↳ but F-test itself does not have power to detect non-linearity

3. If reject $H_0 : \beta_1 = 0$, may interpret the following:

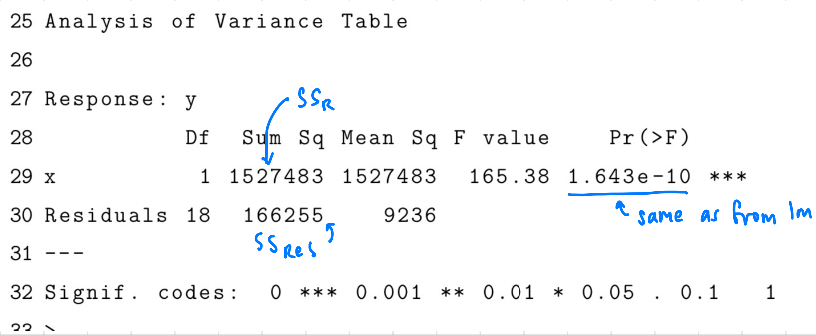
a) this does not mean that SLR is correct, only that the latter predicts better than the intercept-only model.

b) SLR might be wrong, with every single one of assumptions violated, and yet better than the intercept-only model.

F-test ≡ t-test for SLR (proof in handout)

Oct 29

ANOVA table in R



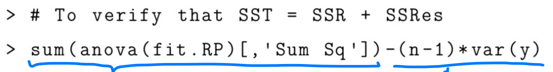
```
Analysis of Variance Table

Response: y
           Df  Sum Sq Mean Sq F value    Pr(>F)
x           1 1527483 1527483  165.38 1.643e-10 ***
Residuals  18  166255    9236
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

SS_R

SS_{Res}

same as from lm

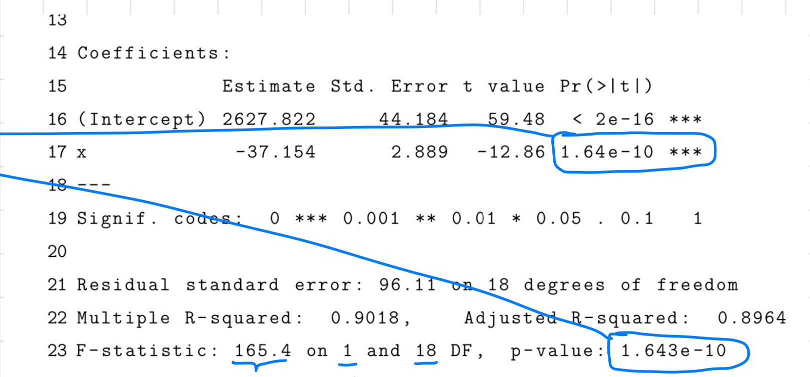
verify $SS_T = SS_R + SS_{Res}$



```
> # To verify that SST = SSR + SSRes
> sum(anova(fit.RP)[,'Sum Sq'])-(n-1)*var(y)
```

$SS_{Res} + SS_R$ should be 0.

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



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2627.822     44.184   59.48  < 2e-16 ***
x            -37.154      2.889  -12.86 1.64e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 96.11 on 18 degrees of freedom
Multiple R-squared:  0.9018,	Adjusted R-squared:  0.8964
F-statistic: 165.4 on 1 and 18 DF,  p-value: 1.643e-10
```

notice that these 2 values are the same! F ≡ t test for SLR

not necessarily true for multiple lin reg.

$T_1^2 = F_0$

$p-1$

$n-p$

Prediction Inference

At an arbitrary value $X_1 = x_{01}$ (not necessarily contained in the training data), we predict that on average Y will be: $\hat{y} = \hat{m}(x_{01}) = \hat{\beta}_0 + \hat{\beta}_1 x_{01}$

$\uparrow$ we dont actually know what y will be! only know what it looks like on average.

$\nwarrow \approx \mathbb{E}(Y \mid X_1 = x_{01})$

The point prediction $\hat{y}$ is called the fitted value of the regression at $X_1 = x_{01}$.
Thus, $\hat{m}(x_{01})$ is an estimate of $\mathbb{E}(Y \mid X_1 = x_{01})$.
↖ random, function of data ↖ deterministic qty

$\hat{m}(x_{01})$ inherits randomness from the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, which in turn inherit theirs from Y.

Confidence interval for conditional mean

A $100(1-\alpha)\%$ CI for the conditional mean at the point $X_1 = x_{01}$ is:

$$CI(m(x_{01})) = [\hat{m}(x_{01}) - k \cdot ese(\hat{m}(x_{01})), \hat{m}(x_{01}) + k \cdot ese(\hat{m}(x_{01}))]$$

Recall that $k = t_{\frac{\alpha}{2}, n-2}$, it is the $[\frac{\alpha}{2} + (1-\alpha)]$ quantile of t
→ in R: $k = qt(1 - \frac{\alpha}{2}, df = n-2)$

How to calculate?

$$\hat{m}(x_{01}) \pm k \cdot ese(\hat{m}(x_{01})) = \hat{m}(x_{01}) \pm k \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_{01} - \bar{x})^2}{S_{xx}}\right)}$$

→ proof in handout about CI theory

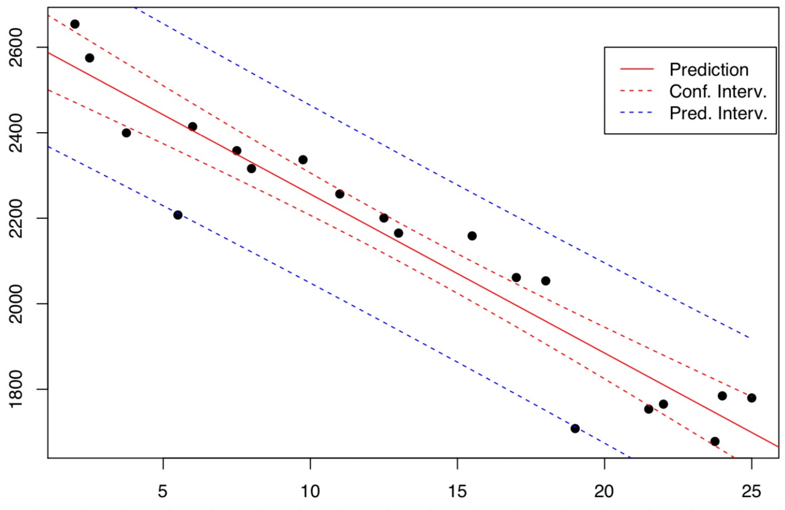
$\mathbb{E}(\hat{m}(x_{01})) = \beta_0 + \beta_1 x_{01}$ (unbiased)

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

The CI for cond. mean characterizes uncertainty in the prediction.

On the other hand, the prediction interval characterizes uncertainty in the data.

Prediction interval · **Nov 3**



CI is narrower than the pred. int.

Using GN-SLR, we know that at $X_1 = x_{01}$ the future observation of Y_0 is:

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

$x_{11}, x_{12}, \dots x_{1n}$ is the training data.
x_{01} could be any one of x_{1i}, but could also be any arbitrary value.

Now we could construct a CI for Y_0 instead of $m(x_{01})$, we call it the pred. interval for the future observation of Y_0 corresponding to $X_1 = x_{01}$.

$$PI(Y_0) = [\hat{Y}_0 - k \cdot ese(\hat{Y}_0) \, , \, \hat{Y}_0 + k \cdot ese(\hat{Y}_0)]$$

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

Note:

- A $100(1-\alpha)\%$ CI on conditional mean $m(x_{01})$ is an interval $[a, b]$ where
  $P(a \le m(x_{01}) \le b) = 1 - \alpha$
  ↑ true reg function: $\beta_0 + \beta_1 x_{01}$

- A $100(1-\alpha)\%$ PI for a future observation $Y_0 = m(x_{01}) + \varepsilon$ is an interval $[a', b']$:
  $P(a' \le Y_0 \le b') = 1 - \alpha$.

Asymptotic behavior when $n \to \infty$

For $CI(m(x_{01}))$:

$$\hat{m}(x_{01}) \pm k \cdot \sqrt{\hat{\sigma}^2 (\frac{1}{n} + \frac{(x_{01} - \bar{x}_1)^2}{S_{xx}})} \xrightarrow{n \to \infty} m(x_{01}) = \beta_0 + \beta_1 x_{01}$$

- $\hat{m}(x_{01}) \xrightarrow{n \to \infty} m(x_{01})$ since $\hat{\beta}_0, \hat{\beta}_1 \to \beta_0, \beta_1$ as $n \to \infty$
- $\frac{(x_{01} - \bar{x}_1)^2}{S_{xx}} = \frac{(x_{01} - \bar{x}_1)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \xrightarrow{n \to \infty} 0$

Thus $CI(m(x_{01})) \xrightarrow{n \to \infty} m(x_{01})$, a single point.

For $PI(Y_0)$:

$$\hat{m}(x_{01}) \pm k \cdot \sqrt{\hat{\sigma}^2 (1 + \frac{1}{n} + \frac{(x_{01} - \bar{x}_1)^2}{S_{xx}})} \xrightarrow{n \to \infty} m(x_{01}) \pm k\sigma$$

Thus width of $PI(Y_0) = 2k\sigma$ as $n \to \infty$

$$PI(Y_0) \xrightarrow{n \to \infty} [m(x_{01}) - k\sigma \, , \, m(x_{01}) + k\sigma]$$

Correlation coefficient

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}} = \frac{\sigma_{XY}}{\sigma_X\,\sigma_Y}$$

estimator: $$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2\right]^{\frac{1}{2}}} = \frac{S_{xy}}{\left[S_{xx}\;SS_T\right]^{\frac{1}{2}}}$$

Note that $$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \underbrace{\left(\frac{SS_T}{S_{xx}}\right)^{\frac{1}{2}}}_{\uparrow\ \sqrt{\frac{\text{spread of } Y}{\text{spread of } x}}} r$$

Furthermore, $r^2 = R^2$:

$$\hat{\beta}_1^{\,2}\left(\frac{S_{xx}}{SS_T}\right) = \frac{\hat{\beta}_1 S_{xy}}{SS_T} = \frac{SS_R}{SS_T} = R^2$$

Multiple Linear Regression

(LR with multiple predictors)

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = \underbrace{X}_{[1\;\; X_1\;\; X_2\;\; \cdots\;\; X_k]}\underbrace{\beta}_{\begin{bmatrix}\beta_0\\ \vdots\\ \beta_k\end{bmatrix}}$$

Polynomial regression: $E[Y|X_1] = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \cdots + \beta_k X_1^k$
(single predictor X_1, just transformed)
→ shape of the relationship between X and Y.

Model with interactions: allow for joint effect of predictors.

$$E[Y|X_1] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 = \underbrace{X}_{[1\;\; X_1\;\; X_2\;\; X_1X_2]}\underbrace{\beta}_{\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_{12}\end{bmatrix}}$$

- without interaction: $E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  with ": $E(Y|X) = \beta_0 + \beta_1 X_1 + (\beta_2 + \beta_{12} X_1) X_2$

Model with transformation: $E(Y|X) = \beta_0 + \beta_1 \log X_1 + \beta_2 X_2^2 + \beta_3 e^{X_3}$

In all cases we are using $E(Y|X) = X\beta$, $X \in \mathbb{R}^{p}$ ← # of pred.

MLR assumptions:

1. There are k predictors, $X_1, \cdots, X_k$. No assumptions about their distribution (might be random or non-random).
   We denote X (no subscript) as $[X_1 \;\cdots\; X_k]$

2. Single response Y: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon = X\beta + \varepsilon$

3. Noise $\varepsilon \perp X$: $E(\varepsilon) = E(\varepsilon|X) = 0$, $Var(\varepsilon) = Var(\varepsilon|X) = \sigma^2$
   ε_i's from different observations are independent ($\varepsilon_j \perp \varepsilon_i$)
   $\Rightarrow Var(\varepsilon|X) = \sigma^2 I_n$

If given n observations, $p = k+1$ parameters,

$$\underset{(n\times 1)}{\mathbb{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}} \qquad \underset{(n \times p)}{\mathbb{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{nk} \end{bmatrix}} \qquad \underset{(p\times 1)}{\beta = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix}} \qquad \underset{(n \times 1)}{\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}}$$

NOV 10

$\mathbb{Y} = \mathbb{X}\beta + \varepsilon$ where:

$\mathbb{E}(\varepsilon \mid \mathbb{X}) = 0_n$

Variance-cov matrix:
$$\begin{bmatrix} Var(\varepsilon_1) & Cov(\varepsilon_1, \varepsilon_2 \mid \mathbb{X}) & \cdots & Cov(\varepsilon_1, \varepsilon_n \mid \mathbb{X}) \\ Cov(\varepsilon_2, \varepsilon_1 \mid \mathbb{X}) & Var(\varepsilon_2) & & \\ \vdots & & \ddots & \\ Cov(\varepsilon_n, \varepsilon_1 \mid \mathbb{X}) & & & Var(\varepsilon_n) \end{bmatrix} = \underset{(n \times n)}{\begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix}}$$

$$\Rightarrow \mathbb{E}(\mathbb{Y} \mid \mathbb{X}) = \mathbb{X}\beta_{\leftarrow n \times 1}$$

Parameter Interpretation

σ^2: variance of noise around the true reg function (hyperplane).
$=$ variance of Y that cannot be explained by X.

β_0: expected value of Y when $X_1 \ldots X_k$ are all 0
$\mathbb{E}[Y \mid X_1 = 0, \ldots, X_k = 0] = \beta_0$

β_j for $j = 1, \ldots, k$: $\mathbb{E}(Y \mid X_j = x_j + 1, X_{-j} = x_{-j}) - \mathbb{E}(Y \mid X_j = x_j, X_{-j} = x_{-j})$
$= [\beta_0 + \beta_j(x_j + 1) + \cdots] - [\beta_0 + \beta_j x_j + \cdots]$
$= \beta_j$

Interpretation: If we select 2 sets of cases from the distribution of the data where X_j differs by 1, we expect Y to differ by β_j on average when all remaining predictors X_{-j} are held constant.

Least Squares for MLR

Observed data: $(y_1, X_1), (y_2, X_2), \ldots, (y_n, X_n)$

$\nwarrow X_1 = [1 \;\; x_{11} \;\; x_{12} \;\; \cdots \;\; x_{1k}]$
$\vdots$
$X_n = [1 \;\; x_{n1} \;\; x_{n2} \;\; \cdots \;\; x_{nk}]$

Goal: estimate $B = (\beta_0 \; \beta_1 \; \cdots \; \beta_k)^T$ in $\mathbb{E}(\mathbb{Y} \mid \mathbb{X}) = \mathbb{X}\beta$
→ minimize sums of squared residuals:

$$S(\beta) = \sum_{i=1}^n (y_i - X_i\beta)^2 = (\mathbb{Y} - \mathbb{X}\beta)^T(\mathbb{Y} - \mathbb{X}\beta) = \underbrace{\|\mathbb{Y} - \mathbb{X}\beta\|_2^2}_{n \times 1}$$

$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \|\mathbb{Y} - \mathbb{X}\beta\|_2^2$, solve by taking derivative wrt $\beta = (\beta_0, \beta_1, \ldots \beta_k)$

$$\frac{\partial S(\beta)}{\partial \beta_j} = -2\sum_{i=1}^n x_{ij}(y_i - X_i\beta) := 0 \quad \text{for } j = 1, \ldots, k$$

In matrix form: $\nabla_\beta S(\beta) = -2\,\mathbb{X}^T(\mathbb{Y} - \mathbb{X}\beta) = 0_p$

⇔ solving the equation $\mathbb{X}^T\mathbb{Y} - \mathbb{X}^T\mathbb{X}\beta = 0$

$$\mathbb{X}^T\mathbb{X}\hat{\beta} = \mathbb{X}^T\mathbb{Y}$$

$$\Rightarrow \hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

↑ but is this invertible?

Some conditions for $(\mathbb{X}^T\mathbb{X})$ invertible: $\mathbb{X} \in \mathbb{R}^{n \times p}$.

- There must be more data than the number of parameters, $n \geq p$.
- In $\mathbb{X}$, columns are not linearly dependent

The fitted value of the reg model at $x_i = (1\ x_{i1}\ x_{i2}\ \dots\ x_{ik})$, $i = 1 \dots n$

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$$

For $\mathbb{X}$: $\hat{\mathbb{Y}} = \mathbb{X}\hat{\beta} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$

Statistical properties of $\hat{\beta}$

- $E(\hat{\beta}) = E(\hat{\beta}|\mathbb{X}) = \beta$ unbiased
- $\underset{(p \times p)}{Var(\hat{\beta})} = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1} = \sigma^2 C$

NOV 12

Estimator of σ^2

$$Var(\varepsilon) = \mathbb{E}(\varepsilon^2) - (\mathbb{E}(\varepsilon))^2 = \mathbb{E}(\varepsilon^2) \approx \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$$

$$= \frac{SS_{Res}}{n-p}$$

$$= \frac{(\mathbb{Y}-\hat{\mathbb{Y}})^T(\mathbb{Y}-\hat{\mathbb{Y}})}{n-p}$$

$$= \frac{(\mathbb{Y}-\mathbb{X}\hat{\beta})^T(\mathbb{Y}-\mathbb{X}\hat{\beta})}{n-p}$$

$\therefore \hat{\sigma}^2 = \dfrac{\|\mathbb{Y}-\mathbb{X}\hat{\beta}\|_2^2}{n-p}$, $p = k+1$, k = # of predictors.

↖ Note that $n \approx n-p$ when p is fixed and n is large

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

t-test in MLR

Recall that for SLR, we required Gaussian-Noise assumption to construct t tests.

Therefore we further assume:

4. $\varepsilon \sim N(0_n, \sigma^2 \mathbb{I}_n)$ $\quad \varepsilon = \begin{bmatrix}\varepsilon_1 \\ \vdots \\ \varepsilon_n\end{bmatrix}$

5. $\varepsilon \perp \mathbb{X}$. It follows that conditional on $\mathbb{X}$, $\mathbb{Y}$ has a multivariate Gaussian distribution: $\mathbb{Y}|\mathbb{X} \sim N(\mathbb{X}\beta, \sigma^2 \mathbb{I}_n)$

We can show that under GN-MLR, $\hat{\beta} \sim N(\beta, \underset{\overset{\uparrow}{\text{unknown, use } \hat{\sigma}^2}}{\sigma^2} (\mathbb{X}^T\mathbb{X})^{-1})$

It follows that $\underset{\text{j}^{\text{th}} \text{ entry of } \hat{\beta}}{\hat{\beta}_j} \sim N(\beta_j, \sigma^2[(\mathbb{X}^T\mathbb{X})^{-1}]_{jj})$.

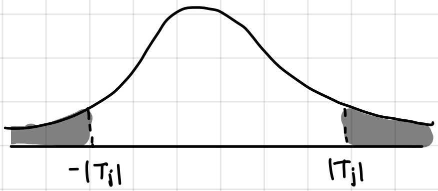
$H_0 : \beta_j = 0 \qquad j = 0, 1, \ldots, k$

$H_1 : \beta_j \neq 0$

If we reject H_0, it indicates that the predictor X_j is likely to be statistically detectable / significant in the model.

$$T_j = \frac{\hat{\beta}_j - 0}{ese(\hat{\beta}_j)}, \quad ese(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 \cdot C_{jj}} = \sqrt{\underset{\nwarrow = \frac{SS_{Res}}{n-p}}{\hat{\sigma}^2} [(\mathbb{X}^T\mathbb{X})^{-1}]_{jj}}$$

Under GN-MLR and under H_0, $T_j \sim t_{n-p}$.



$P(T < -|T_j| \text{ or } T > |T_j|) =$ p-value

Reject H_0 if $p < \alpha$

We reject H_0 if $|T_j| > k = t_{\frac{\alpha}{2}, n-p} \longleftarrow (1-\frac{\alpha}{2})100\%$ quantile in t_{n-p} dist.

F-test in MLR

Consider the regression model $\mathbb{Y} = \mathbb{X}\beta + \varepsilon$, $\beta = (\beta_0, \beta_1, \ldots, \beta_4)$

Test: $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$

H_1 : at least 1 is not 0.

$\Leftrightarrow H_0 : Y = \beta_0 + \beta_1 X_1$ (reduced model)

$H_1 : Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_4 X_4$ (full model)

Nov 17

Generally, $\mathbb{Y} = \mathbb{X}\beta + \varepsilon = \underset{\text{partition}}{[\mathbb{X}_{(1)} \ \mathbb{X}_{(2)}]} \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \end{bmatrix} + \varepsilon$

$\mathbb{X}_{(1)} : (n \times (p-r)) \qquad \beta_{(1)} : ((p-r) \times 1)$

$\mathbb{X}_{(2)} : (n \times r) \qquad \beta_{(2)} : (r \times 1)$

$$\left.\begin{array}{l} H_0 : \beta_{(2)} = 0_r \\ H_1 : \beta_{(2)} \neq 0_r \end{array}\right\} \text{equivalent to} \quad \begin{array}{l} H_0 : \mathbb{Y} = X_{(1)}\beta_{(1)} + \varepsilon \\ H_1 : \mathbb{Y} = X_{(1)}\beta_{(1)} + X_{(2)}\beta_{(2)} + \varepsilon \end{array}$$

The null hypothesis $\beta_{(2)} = 0$ can be tested by F-statistic:

$$F_0 = \frac{\overline{SS}_R(\beta_{(2)} | \beta_{(1)}) / r}{\underbrace{SS_{Res}(\beta) / (n-p)}_{MS_{Res}}}, \quad \overline{SS}_R(\beta_{(2)} | \beta_{(1)}) = \overline{SS}_R(\beta) - \overline{SS}_R(\beta_{(1)})$$

$$\overline{SS}_R(\beta) = \hat{\beta}^T X^T Y \qquad \overline{SS}_R(\beta_{(1)}) = \hat{\beta}_{(1)}^T \mathbb{X}_{(1)}^T Y$$

Recall that $\hat{\beta} = (X^TX)^{-1}X^TY$, $\hat{\beta}_{(1)} = (X_{(1)}^TX_{(1)})^{-1}X_{(1)}^TY$

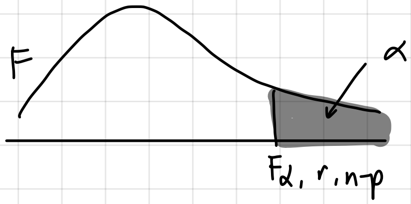
$$MS_{Res} = \frac{SS_{Res}}{n-p} = \frac{Y^TY - \hat{\beta}^TX^TY}{n-p}$$

$\overline{SS}_R(\beta_{(2)} | \beta_{(1)})$ has degrees of freedom $p - (p-r) = r$
It is the extra sums of squares due to $\beta_{(2)}$ when $\beta_{(1)}$ is included in the model.
i.e. extra contribution in $\overline{SS}_R$ due to predictors in $X_{(2)}$.

We can show that under GN-MLR assumptions and under $H_0: \beta_{(2)} = 0$, the random qty F_0 follows the F distribution with degrees of freedom r, $n-p$.

$$F_0 \sim F_{r,\, n-p}$$

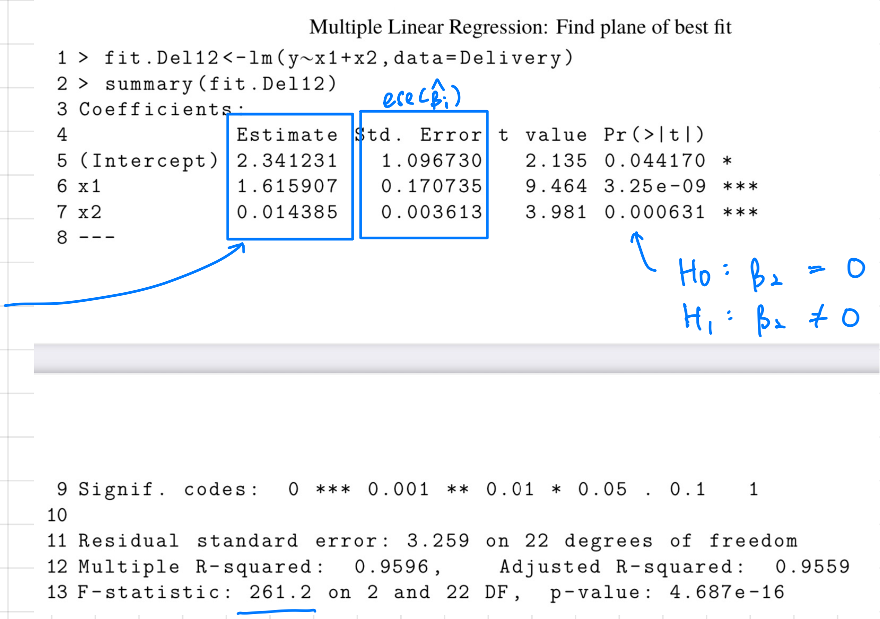
We reject H_0 if: $F_0 > F_{\alpha,\, r,\, n-p}$
where $F_{\alpha,r,n-p}$ is $(1-\alpha)100\%$ quantile of the F distribution with r, $n-p$ d.f.



equivalently,
p-value = $P(F > F_0)$
Reject Ho when p-value $< \alpha$.

Conclusion: If we reject Ho, this means that at least 1 of the parameters in $\beta_{(2)}$ is not zero $\Rightarrow$ at least 1 of the predictors in $X_{(2)}$ is not zero.

(B) Multiple Linear Regression



Multiple Linear Regression: Find plane of best fit

```
> fit.Del12<-lm(y~x1+x2,data=Delivery)
> summary(fit.Del12)
Coefficients:

(Intercept) 2.341231   1.096730    2.135 0.044170 *
x1          1.615907   0.170735    9.464 3.25e-09 ***
x2          0.014385   0.003613    3.981 0.000631 ***
---

Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

```
            Estimate  Std. Error t value Pr(>|t|)
```

$ese(\hat{\beta}_i)$

$\hat{\beta} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2]$

$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

F test: $H_0: y = \beta_0 + \varepsilon$ (intercept only) } $H_0: \beta_{(2)} = 0$ $\quad \beta_{(2)} = (\beta_1, \beta_2)$
$H_1: y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ } $H_1: \beta_{(2)} \neq 0$

An equivalent formula for F-statistic

Nov 19

Recall that F statistic:

$$F_0 = \frac{\overline{SS}_R(\beta_{(2)} \mid \beta_{(1)})/r}{SS_{Res}(\beta)/(n-p)}$$

$$= \frac{(\overline{SS}_R(\beta) - \overline{SS}_R(\beta_{(1)}))/r}{SS_{Res}(\beta)/(n-p)}$$

Equivalently, $F_0 = \dfrac{[SS_{Res}(\beta_{(1)}) - SS_{Res}(\beta)]/r}{SS_{Res}(\beta)\ /\ (n-p)}$ ← this calculation is used in R

→ Proof: Originally we have $SS_T = SS_{Res} + SS_R$.

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

Now we consider an alternate form:

$$\overline{SS}_T = SS_{Res} + \overline{SS}_R$$

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

$$(Y - \hat{Y})^T (X\hat{\beta}) = \underbrace{(Y - \hat{Y})^T X}_{\text{normal eqn}, \ = 0} (X^TX)^{-1} X^T Y$$

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

$$= \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n \hat{y}_i^2 + \underbrace{2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i)}_{=0}$$

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

Specifically in MLR: $\overline{SS}_R = \sum_{i=1}^n \hat{y}_i^2 = (X\hat{\beta})^T(X\hat{\beta})$

$$= \hat{\beta}^T X^T X (X^TX)^{-1} X^T Y$$

$$= \hat{\beta}^T X^T Y$$

$$\overline{SS}_T = SS_{Res}(\beta) + \overline{SS}_R(\beta) = SS_{Res}(\beta_{(1)}) + \overline{SS}_R(\beta_{(1)})$$

$$\Rightarrow \overline{SS}_R(\beta) - \overline{SS}_R(\beta_{(1)}) = SS_{Res}(\beta_{(1)}) - SS_{Res}(\beta) \quad ▨$$

CI in MLR

1. CI for β
2. CI for conditional mean $E(Y \mid X)$
3. Prediction interval for Y^0 (new observation)

see handout

Model diagnostic

We hope to cheek the adequacy of our model. The residuals are:

$$e_i = y_i - \hat{y}_i$$
$$= y_i - x_i\hat{\beta} \quad , \quad i = 1, \dots, n$$

In matrix format: $e = (e_1\ e_2\ \cdots\ e_n)^T$

$$e = \mathbb{Y} - \hat{\mathbb{Y}}$$
$$= \mathbb{Y} - \mathbb{X}\hat{\beta}$$
$$= \mathbb{Y} - \underbrace{\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T}_{H}\mathbb{Y} = (I_n - H)\mathbb{Y}$$

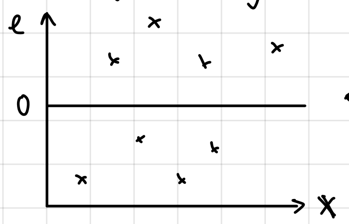
H is symmetric: $H^T = H$
idempotent: $HH = H$

NOV 24

1. $\mathbb{E}(e|\mathbb{X}) = 0_n$

since
$$\mathbb{E}(e|\mathbb{X}) = \mathbb{E}(\mathbb{Y} - \hat{\mathbb{Y}}|\mathbb{X})$$
$$= \mathbb{E}(\mathbb{Y}|\mathbb{X}) - \mathbb{E}(\mathbb{X}\hat{\beta}|\mathbb{X})$$
$$= \mathbb{X}\beta - \mathbb{X}\beta \quad \nwarrow \text{unbiased}$$
$$= 0_n$$

We plot e against any predictor X_j or the linear combination of any predictors:



constant mean function

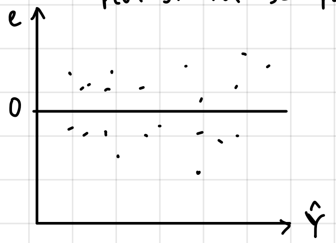
2. We already know that $\mathbb{E}(e|\hat{\mathbb{Y}}) = 0$

Since
$$\mathbb{E}(e|\hat{\mathbb{Y}}) = \mathbb{E}(e|\mathbb{X}\hat{\beta})$$
$$= \mathbb{E}(e|\mathbb{X}) = 0$$

Also, we have $Cov(e, \hat{\mathbb{Y}}) = 0 \in \mathbb{R}^{n\times n}$

since
$$Cov(e, \hat{\mathbb{Y}}) = Cov((I_n - H)\mathbb{Y}, H\mathbb{Y})$$
$$= (I_n - H)H \cdot Var(\mathbb{Y})$$
$$= \sigma^2(I_n - H)H$$
$$= \sigma^2(H - HH)$$
$$= 0 \quad \nwarrow H \text{ idempotent}$$

plot should be patternless



no linear trend

zero mean function.

3. Variation of e: The variance of residual e $Var(e|\mathbb{X}) = \sigma^2(I_n - H)$

since
$$var(e|\mathbb{X}) = Var((I_n - H)\mathbb{Y}|\mathbb{X})$$
$$= (I_n - H)^T(I_n - H)\ Var(\mathbb{Y}|\mathbb{X})$$
$$= (I_n - H)\sigma^2 \quad \text{which is not a diagonal matrix}$$

$$[var(e|\mathbb{X})]_{ii} = (1 - h_{ii})\sigma^2 = var(e_i|\mathbb{X})$$

$\nwarrow$ the i^{th} diagonal entry of H

$$[var(e|\mathbb{X})]_{ij} = -\sigma^2 h_{ij} \neq 0$$
$$= cov(e_i, e_j|\mathbb{X})$$

$\leftarrow$ e_i, e_j are correlated unlike noise terms $\varepsilon_1 \dots \varepsilon_n$ which are independent.

In SLR case, $Var(e_i|X) = \sigma^2(1-h_{ii}) = \sigma^2\left(1-\frac{1}{n}-\underset{\uparrow\ \text{small!}}{\frac{(x_i-\bar{x})^2}{S_{xx}}}\right) \approx \underbrace{\sigma^2\left(1-\frac{1}{n}\right)}_{\text{a constant}}$

$$\frac{(x_i-\bar{x})^2}{S_{xx}} = \frac{(x_i-\bar{x})^2}{(x_1-\bar{x})^2+\dots+(x_n-\bar{x})^2}$$

Therefore the points near the center $\bar{x}$ have larger variance for e_i than e_i at more remote locations.

We should expect to see $Var(e_i|X) \approx \sigma^2(1-\frac{1}{n})$ which is almost constant

The difference in $\frac{(x_i-\bar{x})^2}{S_{xx}}$ gets less pronounced for a larger n.

4. For GN model (SLR and MLR)
If $Y|X \sim N$, then $e = (e_1, \dots, e_n)^T$ is also Gaussian.
$e_i \sim N(0, \sigma^2)$ for $i = 1, \dots, n$
↖ approximately true for large n.

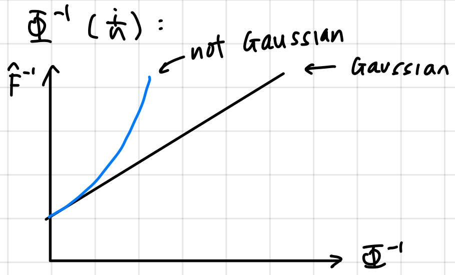
Quantile-Quantile (Q-Q) plot: make a histogram of e and compare to Gaussian
For a continuous dist, CDF: $F(z) = P(Z \le z)$ has an inverse function:
$F^{-1}(p) = z$ s.t. $p = P(Z \le z)$
For a Gaussian $Z \sim N(\mu, \sigma^2)$, quantile function: $F^{-1}(p) = \sigma\Phi^{-1}(p) + \mu$
where Φ is the quantile function for $N(0,1)$

If we plot F^{-1} against Φ^{-1} we should get a straight line.
In practice, we replace F^{-1} by the quantile function of e, $\hat{F}^{-1}$
We estimate $\hat{F}^{-1}$ by using residuals $e_1 \dots e_n$
Arrange $e_1 \dots e_n$ in increasing order: $e_{(1)}, e_{(2)}, \dots, e_{(n)}$
$e_{(i)}$ is greater than $\frac{i}{n}$ of the residuals.

$$\hat{F}^{-1}\left(\frac{i}{n}\right) = e_{(i)}$$

Now we can plot $\hat{F}^{-1}(\frac{i}{n}) = e_{(i)}$ against $\Phi^{-1}(\frac{i}{n})$:

$$\begin{aligned} \hat{F}^{-1}\left(\tfrac{1}{n}\right) &\longrightarrow \Phi^{-1}\left(\tfrac{1}{n}\right) \\ \hat{F}^{-1}\left(\tfrac{2}{n}\right) &\longrightarrow \Phi^{-1}\left(\tfrac{2}{n}\right) \\ &\vdots \\ \hat{F}^{-1}\left(\tfrac{n}{n}\right) &\longrightarrow \Phi^{-1}\left(\tfrac{n}{n}\right) \end{aligned}$$



Test zero mean, constant variance → residual plot

Nov 26

Test Gaussian assumptions of noise → plot histogram of residuals,
Q-Q plot

Test prediction performance of the model → generalization error

Generalization error

$X = [1 \quad X_1 \quad \cdots \quad X_k]^T$

Response variable Y, predictors $X \in \mathbb{R}^p$.

We build a predictive model $\hat{f}(X)$ from the training data T.

- in MLR setting:

$$T = (y_1, x_1) \ \ldots \ (y_n, x_n) = (\mathbb{Y}, \mathbb{X})$$

$$\hat{f}(X) = X^T \hat{\beta} = X^T (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$$

To evaluate $\hat{f}(X)$, introduce a loss function:

$$L(Y, \hat{f}(X)) = (Y - \hat{f}(X))^2 \quad \text{or} \quad |\underset{\text{target}}{Y} - \underset{\text{prediction}}{\hat{f}(X)}|$$

Generalization error: $Y, X \sim$ test data

$$\text{Error}_T = \mathbb{E}_{Y,X}(L(Y - \hat{f}(X)) \mid T)$$

$$\approx \frac{1}{n'} \sum_{i=1}^{n'} L(y_i, \hat{f}(x_i))$$

$(y_1, x_1) \ldots (y_n, x_n)$ are from test data, n' is the size of test data.

Generalization error vs training error:

- training error:

training data $T = \{(y_i, x_i)\}_{i=1}^n$

$$err = \mathbb{E}_{\underset{\text{from } T}{Y,X}}(L(Y, \hat{f}(X)) \approx \frac{1}{n} \sum_{i=1}^n L(y_i, \hat{f}(x_i))$$

in MLR, $err = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 = \frac{1}{n} (\mathbb{Y} - \mathbb{X}\hat{\beta})^T (\mathbb{Y} - \mathbb{X}\hat{\beta})$

- In general, generalization error > training error.

Assume data generated from the model $Y = X\beta + \varepsilon$. For simplicity, fix X.

Test: $Y' = X\beta + \varepsilon'$

Generalization error (test error): $\mathbb{E}_{Y'}[\frac{1}{n} \sum_{i=1}^n (Y_i' - \hat{Y}_i)^2]$

Training error: $\mathbb{E}_Y[\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2]$

We can show that:

$$\underset{\text{(test error)}}{\mathbb{E}_{Y'}[\frac{1}{n} \sum_{i=1}^n (Y_i' - \hat{Y}_i)^2]} = \underset{\text{(training error)}}{\mathbb{E}_Y[\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2]} + \frac{2\overset{p = k+1 \text{ (total \# of variables)}}{p}}{\underset{\text{\# of samples in data}}{n}} \underset{\varepsilon, \varepsilon' \sim N(0, \sigma^2)}{\sigma^2}$$

See handout for proof

Approximation of generalization error:

$$\frac{1}{n}\sum_{i=1}^{n} (Y_i' - \hat{Y}_i)^2 \approx \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \frac{2p}{n}\sigma^2$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \frac{2p}{n}\hat{\sigma}^2$$

Mallow's Cp statistic:

$$C_p = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \frac{2p}{n}\hat{\sigma}^2$$

e.g. 3 pred. models $\hat{f}_1, \hat{f}_2, \hat{f}_3$.

$\mathbb{E}(Y|X) = \hat{\beta}_0 + \hat{\beta}_1 X_1$

$\mathbb{E}(Y|X) = \hat{\beta}_0$

$\mathbb{E}(Y|X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$

Usually $\hat{\sigma}^2$ is obtained from the largest model i.e. $\hat{f}_3$

Whichever model gives the smallest C_p value is the best in pred. performance

Adjusted R^2:

Recall that adjusted R^2 is $R^2_{adj} = 1 - \dfrac{\overset{err}{\frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)} \cdot \frac{n}{n-p}}{\underbrace{\frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2}_{\frac{1}{n} S_{yy}}}$

maximization of $R^2_{adj} \iff$ minimize $err \cdot \frac{n}{n-p}$

$$err \cdot \frac{n}{n-p} = err \cdot \frac{1}{1-\frac{p}{n}}$$

$$\approx err\left(1 + \frac{p}{n}\right) = err + \frac{p}{n} \cdot err$$

When $n \to \infty$, $err \to \sigma^2$

Thus $err \cdot \frac{n}{n-p} \to err + \frac{p}{n}\sigma^2$ as $n \to \infty$

AIC:

$$AIC = -\frac{2}{n} \cdot \overset{-\frac{err}{2\sigma^2}}{\text{loglikelihood}} + \frac{2p}{n} \quad \text{(general)}$$

in MLR: $AIC = err + \frac{2p}{n}\hat{\sigma}^2 = C_p$

BIC:

$$BIC = err + \frac{\log n}{n} p \hat{\sigma}^2$$

Transformation

Dec 1

Transformations to linearize the models: If we detect non-linearity in the scatterplots or residual plots, in some case a nonlinear function can be linearized using a suitable transformation.

e.g. $\mathbb{E}(Y|X) = \beta_0 + \beta_1 e^{-X}$

$\mathbb{E}(Y|X) = \beta_0 + \beta_1 (\frac{1}{X})$

$\mathbb{E}(\ln Y|X) = \beta_0 + \beta_1 X_1$

$\mathbb{E}(Y|X) = \beta_0 + \beta_1 \ln X_1$

$\mathbb{E}(\ln Y|X) = \beta_0 + \beta_1 \ln X_1$

e.g. Cobb-Douglas production function: $\underset{\substack{\uparrow \\ \text{Output}}}{O_i} = e^{\beta_0} \underset{\substack{\uparrow \\ \text{labor input}}}{L_i}^{\beta_1} \underset{\substack{\uparrow \\ \text{capital input}}}{C_i}^{\beta_2} \underset{\substack{\uparrow \\ \text{noise}}}{u_i}$

Take log : $\underbrace{\log O_i}_{Y_i} = \beta_0 + \beta_1 \underbrace{\log L_i}_{X_{i1}} + \beta_2 \underbrace{\log C_i}_{X_{i2}} + \underbrace{\log u_i}_{\varepsilon_i}$

Transformation requires $\varepsilon = \log u \sim N(0, \sigma^2)$ if we hope to assume GN-SLR.

Polynomial terms

For any non linear function $f(x) = \underbrace{f(x_0)}_{\text{constant}} + \underbrace{\frac{f'(x_0)}{1!}(x - x_0)}_{\text{(1st order)}} + \underbrace{\frac{f''(x_0)}{2!}(x - x_0)^2}_{\text{(2nd order)}} + \cdots$

we hope to approximate the relationship

$$E(Y|X) = f(x)$$
$$\approx \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \dots + \beta_k X_1^k$$

The more complicated the model (more terms), the more difficult it is to estimate with the same amount of data.

→ the more data you have, the more you can "afford" a more complicated model without overfitting

Factor predictors (for categorical)

A factor predictor is a predictor that takes a discrete set of values on a nominal scale (i.e. non-numerical):

e.g. $X_1 = \{ \text{Drug A}, B, C \}$
$X_1 = \{ \text{male}, \text{female} \}$

Consider the case where X_1 takes M values. We introduce a dummy variable:

$$X_1^{(m)} = \begin{cases} 1 & \text{if } X_1 = m \\ 0 & \text{otherwise} \end{cases}$$

A categorical variable with M unique categories can be represented by M-1 dummies.

For example X_1 has M categories. The corresponding model:

$$E(Y|X) = \beta_0 + \sum_{m=1}^{M-1} \beta_m X_1^{(m)}$$

→ if (M): $X_1^{(1)} \dots X_1^{(M-1)}$ are all 0.

$$E(Y|X) = \beta_0$$