Linear regression *

Simple regression

- Residual sum of squares (RSS) $= e_1^2 + \dots + e_n^2 = (y_1 - \hat{B}_0 - \hat{B}_1 x_1)^2 + \dots + (y_n - \hat{B}_0 - \hat{B}_1 x_n)^2$

least - square minimize RSS with $\hat{B}_1 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x}_n)(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$ $\quad \hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}$

- $\hat{B}_1$ and $\hat{B}_0$ are unbiased (does not systematically over or under estimate the true parameters).

→ In general sample mean unbiased $\hat{u} = \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ and $SE(\hat{u})^2 = \frac{\sigma^2}{n}$, where σ is the SD of each realization y_i of y. $SE(\hat{u})$ tells us how $\hat{u}$ differs from actual value of u.

$SE(\hat{B}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$, $SE(\hat{B}_1)^2 = \left[\frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]$ where $\sigma^2 = Var(\varepsilon)$

supposing common variance σ^2 and uncorrelated.

- Notice in the formula that $SE(\hat{B}_1)$ is smaller when the x_i are more spread out; intuitively we have more leverage to estimate the slope when this is the case. we also see that $SE(\hat{B}_0)$ would be the same as $SE(\hat{u})$ if $\bar{x}$ were zero (in which case $\hat{B}_0 = \bar{y}$).

- σ unknown ⇒ Residual standard error (RSE) $= \sqrt{RSS/(n-2)}$ ⇒ average amount the response will deviate from the true regression line. ⇒ measure lack of fit of the model.

- A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of parameter. $\hat{B}_1 \pm 2 \cdot SE(\hat{B}_1)$

- H_0: there is no relationship between X and Y : $B_1 = 0$ ⇒ $Y = B_0 + \epsilon$ (null)

  H_a : there is some " " " " " $B_1 \neq 0$ ⇒ $Y = B_0 + B_1 X + \epsilon$ (alternative)

we need to test wether $\hat{B}_1$ is sufficiently far away from 0. How far is enough? (because if $SE(\hat{B}_1)$ is high is has to be farther than if $SE(\hat{B}_1)$ is low). so we use

t-statistic ⇒ $t = \dfrac{\hat{B}_1 - 0}{SE(\hat{B}_1)}$ (the number of SD that $\hat{B}_1$ is from 0). with $(n-2)$ degree freedom. (same as gaussian for $n \geq 30$)

we need to compute the probability of observing any number equal to $|t|$ or larger, assuming $\hat{B}_1 = 0$ ⇒ this is the p value. Small p - value (< 5% or 1%) means it is unlikely to observe association between predicator and response due to chance. ⇒ reject null hypothesis

Accuracy of the model (once rejecting the null)

RSE but also R^2 = the proportion of variance explained ($0 \leq R^2 \leq 1$) $R^2 = \dfrac{TSS - RSS}{TSS} = 1 - \dfrac{RSS}{TSS}$

where $TSS = \sum (y_i - \bar{y})^2$ is total sum of squares. which is the variability inherent of the response before the regression is performed).

(TSS - RSS) amount of variability explained by the regression). R^2 measures the proportion of variability in Y that can be explained using X. $R^2 = 1$ good. but hard to dermine the righ R^2 = domain knowledge).

Multiple linear regression

- $H_0: B_1 = B_2 = \cdots = B_p = 0$ H_a: at least one B_j is non zero. This hypothesis is performed using F statistic $F = \dfrac{(TSS-RSS)/p}{RSS/(n-p-1)}$. if the linear assumptions are correct $E\{RSS/(n-p-1)\} = \sigma^2$ and provided H_0 is true $E\{(TSS-RSS)/p\} = \sigma^2$. => iff no relationship F is close to 1, ~~[illegible]~~
  " $> \sigma^2$ iff relationship and $F > 1$

How large does F need to be? if n is big small deviation is strong evidence against H_0. if n is small we need a strong deviation. But in all case, according to n and p, we look at the p-value of F. (because follow f-distribution).

- test a particular subset of q of the coefficients are zero. we fit a new model without the q and get RSS_0 then $F = \dfrac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$

- In multiple regression, we get t-statistic and p-value for each individual (wether it is related to the response). It is equivalent of the F-statistic that omits that single variable, leaving all the others in.
- It seems that if any one of the individual p-value is small, at least one is related to the predictors, but this is flawed. => i.e if $p = 100$ => $B_1 = \ldots B_{100} = 0$, about 5% will be below 0.05 by change. But F-statistic doesn't suffer from that as it ajusts from the number of predictors. Hence if H_0 is true, there is only a 5% change that the F-statistic will result in a p-value below 0.05, regardless the number of predictors. But this approach only when p is small, and small relative to n. otherwise use forward selection or high dimensional method.

- Method in multiple regression → got F_statistic → p-value big (of F) → no relationship
  → p-value of F is small
  - if P is big → Variable selection: we can use Mallow's C_p, AIC, BIC, adjusted R^2 (2.6). But with p predictor we get 2^p differents models so there is easier way.
  - if p is small we can select individually the most promising (but not the best way).

forward selection: begin with a null model of an intercept and no predictor. we then fit p simple linear regression and add to the null model the variable that result in the lowest RSS. we then keep going till a certain threshold.

Backward: we start with all variables, and remove the one with largest p value. And repeat

Mixed: start with nothing, add till p-value for one of the variable ↗, then we remove it. we repeat till small p-value for all.

warning: R^2 always increase (closer to 1) as we add variables, because always reduce the residual sum of squares on the training data. Also $RSE = \sqrt{\dfrac{1}{n-p-1} RSS}$ then we can get ↗ RSE if the decrease in RSS is small relative to the increase in p.