Additional note 3

Lecture 8

1. **Revision questions of L7**

   - If a categorical predictor x has $k = 4$ levels, how many parameters are associated with x in a linear model with an intercept? Without an intercept?

     Answer: There are 3 parameters associated with x in a linear model with an intercept and 4 for model without an intercept (no other categorical x variable). Note that there is no numerical value associated with each class and so there is only a comparison relative to a baseline class in which R uses the first class. So each β_j is a comparison of Y for class j relative to the baseline class 0. So it makes sense to have $k-1$ parameters for each categorical variable of k classes.

   - What is the similarity and difference between a linear model when x is continuous and x is categorical?

     Answer: The parameter estimates of β_j can still be obtained by the equation $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$ with a suitable design matrix $\boldsymbol{X}$ given by the last page of Lecture slide 7 but the difference is the number of parameters being $k-1$ instead of 1.

   - Why should we consider adjusted R^2?

     Answer: We want to adjust or allow for the number of parameter p as some parameters can improve the model fit only marginally but increase the model complexity. So one should penalise the model complexity by using $R^2_a = 1 - \frac{MSR}{MST} = 1 - \frac{SSR/(n-p)}{SST/(n-1)}$ allowing for p and n instead of using $R^2 = 1 - \frac{SSR}{SST}$ without allowing for model complexity. Note that R^2_a can be negative.

   - What does a model $Y \sim x_1 * x_2$ means? Can we drop x_1 from the model?

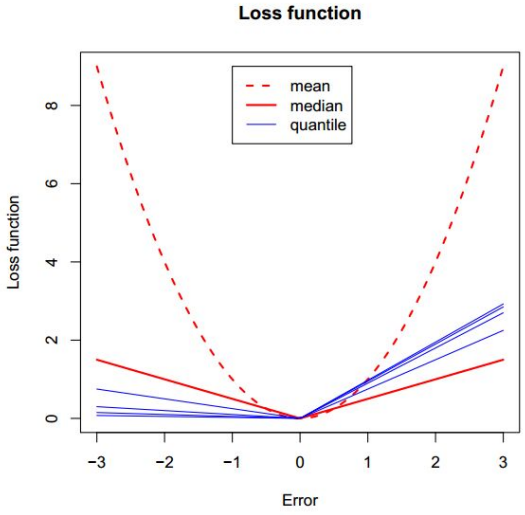
     Answer: This model includes an interaction term between x_1 and x_2 which is equivalent to $Y \sim x_1 + x_2 + x_1 * x_2$. Normally, we consider hierarchical models which contain all lower order terms for all those higher order terms in the model. That means we do not consider $Y \sim x_2 + x_1 * x_2$.

2. **Robust statistics**

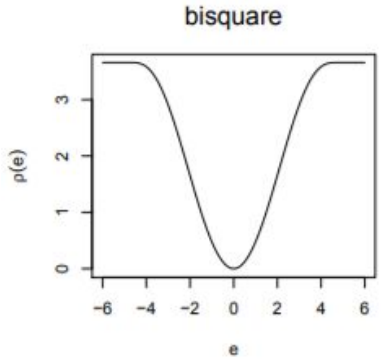
   Recall that SSR (sum of squared residuals) $= \sum_i R_i^2$ uses the square function $y = x^2$ (red dash line in the plot below) which is much higher for large x. Hence the model or $\hat{\boldsymbol{\beta}}$ is more sensitive to outliers as fitting the outliers better can reduce the SSR substantially. Hence it tends to fit the outliers better at the cost of reducing the fit of those middle portion points.

   One of the more sophisticated ways of dealing with outliers in robust statistics is to revise the loss function.

   - We may use say SAR (sum of absolute residuals) $= \sum_i |R_i|$ which contains order 1 (instead of 2) function $y = |x|$ (red solid line) with less weight for large x. Hence the effect of outliers is reduced and hence the model becomes more robust to outliers.



- Alternatively, one may only lower the weight of those outliers which lie outside a certain threshold say $|R_i| > \tau$.



Lecture 9

1. **Revision questions of L8**

   - When will one have high leverage points? How to measure high leverage?

     Answer: For multiple linear regression, points with high leverage are those which are far away from the region or $(p-1)$-dimensional space of $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$. Leverage is measured by $0 < h_{ii} < 1$, the i-diagonal element of matrix $\boldsymbol{H} = \boldsymbol{X}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}$, that is $h_{ii} = \boldsymbol{x}_i^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_i$ and h_{ii} is high if $h_{ii} > 2p/n$. Note that h_{ii} does not depend on Y_i or regression model.

   - For those points with high leverage, when will they have high influence? How to measure the amount of influence and when should it be considered as high?

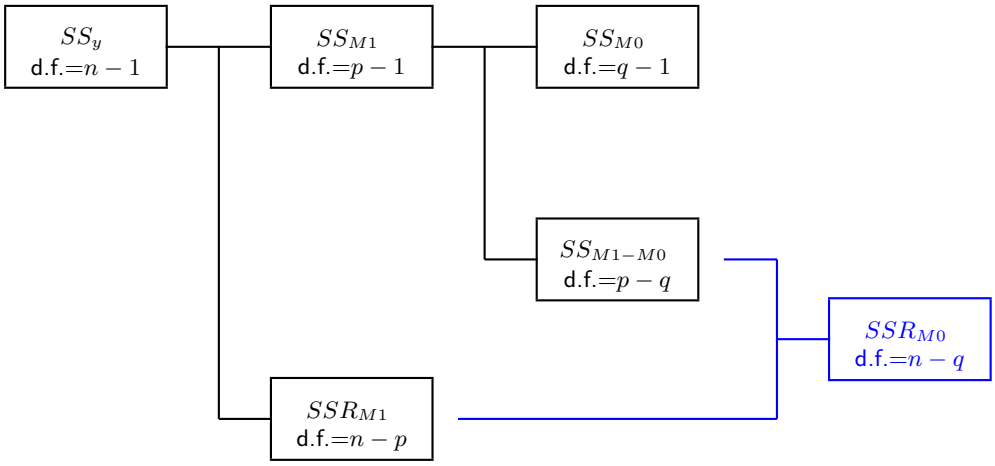
     Answer: For those points with high leverage, they will have high influence if their Y values differ from what is expected under a certain model built by other data. In other word, influence is model dependent which makes sense as it measures the impact of each point on the model. Influence for point i can be measured by Cook's distance which is a standardized sum over observations of squared distance of predicted Y using the whole data set to predicted Y with the i-th observation deleted, that is $D_i = \sum_{j=1}^{n} \frac{(\hat{Y}_j - \hat{Y}_{j[-i]})^2}{p\hat{\sigma}^2} = \frac{R_i^2 h_{ii}}{(1-h_{ii})^2 p\hat{\sigma}^2}$. It is consider large if $D_i > F^{-1}_{p,n-p}(0.5)$.

2. - Note a table of pairwise comparison of 4 models, M0, M1, M2, M3 are given which summarises the all F-tests and other possible F-tests of totally $\binom{4}{2} = 6$ pairs of models. Note that these tests testing between models in H_0 and H_1 should be hierarchical. Hence the F-test between M0 and M3 fails.

- As we are testing a simplified model M0 from M1, M0 should be nested within M1 and the bigger model M1 should be in H_1 whereas the smaller model M0 by setting some constraints (hence reducing the number of parameters) should be in H_0.
- The $\hat{\sigma} = RSS/(n-p)$ in the denominator of the F test statistics is to standardise the difference of RSS between models in H_0 and H_1. We choose to adopt $\hat{\sigma} = RSS_{H_1}(n-p)$ rather than $\hat{\sigma} = RSS_{H_0}(n-q)$ as the bigger model under H_1 is expected to have more explanatory power to estimate σ^2.
- The F-test statistic is given by

$$f_0 = \frac{(RSS_{H_0} - RSS_{H_1})/(p-q)}{RSS_{H_1}/(n-p)}$$

by comparing different sum of squares within the total sum of square SS_y in Y.



Decomposition of total variation in Y in testing H_1: M1 against H_0: M0

Lecture 10

1. **Revision questions of L9**

   - For the Catheter data example with $n = 12$ and the model $L_i = \beta_0 + \beta_1 h_i + \beta_2 w_i + \epsilon_i$, What is the df for the F-test, ie $p-q$ and $n-p$ for the following tests?
   Test 1: $H_0 : \beta_1 = \beta_2 = 0$ vs $H_1 : \beta_1 \neq 0$ or $\beta_2 \neq 0$
   Test 2: $H_0 : \beta_1 = \beta_2 = \beta$ vs $H_1 : \beta_1 \neq \beta$ or $\beta_2 \neq \beta$
   Test 3: $H_0 : \beta_1 = \beta_2 = 0.2$ vs $H_1 : \beta_1 \neq 0.2$ or $\beta_2 \neq 0.2$

     Answer: Test 1: $3-1=2$, $12-3=9$; Test 2: $3-2=1$, $12-3=9$; Test 3: $2-1=1$, $12-2=10$

   - What are the commands to output RSS and $\hat{\sigma}^2$ in the calculation of F-test statistic?

     Answer: If `m1` saves the output of `lm` for model 1 say, `sigma(m1)` gives $\hat{\sigma}$ and `deviance(m1)` gives the RSS. The RSS and df can also be obtained from `anova(m1,m2,..)` where `m1`, `m2` are nested.

2. **Quadratic order of steps in forward or backward selection:**

   Say in a forward selection, the number of steps is $(p-1)+(p-2)+\cdots+k = \frac{(p-1)+k}{2} \times (p-k)$ is a quadratic order function of p where k is the number of x variables in the model to terminate the selection.

---

last adjustments: March 2, 2021 by JC