Questions that people have so far in 617 – Responses to Quiz 2 question

Fall 2022

I tried to organize the questions you have by topic, and then provided answers for each topic. I hope it is helpful to see what questions others in the class have, and to have some answers to those questions.

MATH

To be honest, at first I did not fully understand the implications made in Sheather's book, especially because it has been so long since I looked into these statistical equations. My major issue was that I was not able to make connections between these mathematical terms and real-world business problems. However, as the semester progresses, this point became much more clear, and I am actually re-reading these textbooks thoroughly to solidify my understanding right now. This will take some time since it is quite a bit of reading, but I am making a good progress. I still have a little bit of issue with above mentioned connection (eg. Why would we need to use Hat Matrix in the real-world business context?), but I believe such shortfalls will be resolved soon as I complete my re-reading.

What is the maximum likelihood estimation used for? And is adjusted R square a mainstream judging method on how well the model fits?

> *The mathematics in Sheather's book and in class is a bit hard to get "all at once" so it is great that you are re-reading things when you have the time. The purpose of presenting the math, and having you do some of it, is not to make you an expert at the derivations. Rather, there are two reasons for it:*
>
> 1. *It is good to see some of the "machinery" behind the models we fit, even if we don't have to use that machinery regularly (because R or some other statistical package does the machinery for us).*
> 2. *Having seen some of the machinery, we can have a better understanding of how the model and estimation work, what to expect when things go "right" when fitting a model, and what sorts of things to expect when things go "wrong". We will see more of this when we move into collinarity, variable selection, and other topics.*
>
> *Maximum likelihood estimation is a good example of something that is good to know about, but we would not typically need it very often.*
>
> - *We have already seen in class and in Sheather's book, that estimating β's by maximum likelihood or by least-squares produces the same answer.*
> - *We have also seen that a modified version of maximum likelihood is needed to compute the powers for the Box-Cox transformation*
> - *Later in the semester, maximum likelihood will come up again, when we talk about mixed effects and hierarchical models.*

We will talk more about adjusted R squared when we begin talking about variable selection in the next few lectures. For choosing among models with different numbers of predictors, it is better than regular R squared, but there are still better tools to use.

MATRIX CALCULATIONS

The matrix algebra is something that still mystifies me. I am working on getting a better understanding of it but I find it extremely challenging.

When completing the second homework I became pretty confused on how to write out a hat matrix. I understand that the hat matrix multiplied with y creates yhat, but when I was attempting to fully write out a hat matrix and do the multiplication for myself it became complicated and confused me more. I'm not sure if it's just not something we typically write out since there are so many parts involved in it or if it's not an important thing to communicate.

Early on the matrices were a bit overwhelming because I haven't done linear algebra since Freshmen year. However, HW2 helped me brush up on some matrix properties and ideas (idempotent, symmetry, identity matrix, etc.) so now I feel like I'm on more solid ground. But I still have a bit more ways to go before I feel 100% comfortable with matrices. Generally, I feel that the implementation of the ideas we learn in class into a data analysis helps make the theory more digestible.

I was struggled with Problem 2(b)(iii) and 2(c) in homework2. Other than that, the lecture materials and other homework questions are fine.

Since it has been a long time from my last linear model courses in undergraduate, the matrix and transformation in linear model confuse me a lot.

We've learnt hat matrix in this course, but I don't know the application of hat matrix, what's the meaning of hat matrix?

I don't understand how Hat matrix works in real-world applications since we already have the package. I also don't quite understand how some equations we have in HW#2 but never defined.

I understand all the class materials for far, the only question I have is, why should we prove the matrix in our HW2 questions 1 & 2 because we don't need to use that info in R to run the data.

Something that we covered but still confuses me is how we calculate leverage for the Residuals vs Leverage plots. I understand from a high level what leverage is: data points that have a large affect on the model. However, I still get confused on how you would go about calculating leverage by hand.

If you would like a review of matrix algebra, please take a look at the folder "0 – linear algebra" in the files area of our Canvas site.

The matrix calculations are an example of the mathematics that I am asking you to do, not to become experts, but to better understand a little of the machinery behind applied linear regression, to better understand the answers we get from R and other statistical packages.

- *The hat matrix will return next week when we talk about added-variable plots, and we will see something related to the hat matrix when we examine nonparametric regression and gam's (generalized additive models).*

- *General matrix calculations will appear again when we talk about mixed-effects linear models.*

Again, I will not expect you to become experts in any of these "math" calculations, but by doing some of them I hope that you gain greater understanding of the models. This greater understanding will help you in real-world situations when you have to explain a fitted model or parameter estimates to a client, or you have to explain why a particular fitted model is really no good, compared to a different model.

In the residuals vs leverage plot, the y-axis has the standardized residuals r_i and the x-axis has the leverages h_{ii}. The leverages h_{ii} are the diagonal entries h_{11}, h_{22}, and so forth, of the hat matrix

$$H = \begin{bmatrix} h_{11}\, h_{12} & \dots & h_{1n} \\ h_{21}\, h_{22} & \dots & h_{2n} \\ \vdots & \ddots & \vdots \\ h_{n1}\, h_{n2} & \dots & h_{nn} \end{bmatrix}$$

The standardized residuals are

$$r_i = \frac{\widehat{e_i}}{s\sqrt{1-h_{ii}}} = \frac{y_i - \widehat{y_i}}{s\sqrt{1-h_{ii}}}$$

where s = $\sqrt{RSS/(n-p-1)}$ *. They are not too hard to calculate by hand. Suppose you have a fitted model* `lm.fit`*. Then*

```
X <- model.matrix(lm.fit)
H <- X %*% solve(t(X)%*%X) %*% t(X)
n <- dim(H)[1]
Id <- diag(n)         ## n x n identity matrix
leverages <- diag(H) ## diagonal elements of H
p.plus.1 <- sum(leverages)  ## trace of H
raw.resids <- (Id-H)%*%y
s <- sqrt( t(raw.resids)%*%raw.resids / (n-p.plus.1) )
std.resids <- raw.resids/(s*sqrt(1-leverages))
```

Or you can get them directly from R:

- *The leverages h_{ii} are given by* `leverages <- hatvalues(lm.fit)`
- *The raw residuals $\widehat{e_i}$ are given by* `raw.resids <- residuals(lm.fit)`
- *The standardized residuals r_i are given by* `std.resids <- rstandard(lm.fit)`

We can then get the residuals vs leverage plot with something like

```
plot(std.resids ~ leverages)
```

or "automagically" with

```
plot(lm.fit, which=5)
```

ANOVA TABLE

ANOVA table and the exact meaning of each terms.

The ANOVA table is a way to organize the calculation of an F statistic. In general if we have a "bigger model" M1 and a "smaller model" M0 that is obtained from M1 by setting some of the β coefficients in M1 equal to zero, then the F statistic can be used to test

$$H_0: M0 \text{ holds}, \quad vs \quad H_1: M1 \text{ holds}.$$

The F statistic

$$F = \frac{(RSS_{M0} - RSS_{M1})/(rdf_{M1} - rdf_{M0})}{RSS_{M1}/(rdf_{M1})}$$

is distributed as a F on $(rdf_{M1} - rdf_{M0})$ and (rdf_{M1}) degrees of freedom when H_0 holds (that is, when the smaller model M0 is correct), where rdf is the residual degrees of freedom for each model, so for M1, $rdf_{M1} = n - p_1 - 1$, where p is the number of predictors in M1, and for M0, $rdf_{M0} = n - p_0 - 1$.

The ANOVA table is

Source	df	SS	MS	F
M0	$(rdf_{M0} - rdf_{M1})$	$(RSS_{M0} - RSS_{M1})$	$(RSS_{M0} - RSS_{M1})/(rdf_{M0} - rdf_{M1})$	$F = \frac{(RSS_{M0} - RSS_{M1})/(rdf_{M0} - rdf_{M1})}{RSS_{M1}/rdf_{M1}}$
M1	rdf_{M1}	RSS_{M1}	RSS_{M1}/rdf_{M1}	
Total	rdf_{M0}	RSS_{M0}		

The "total" line isn't so important but the first two lines tell us what goes in the numerator and denominator of the F statistic. These calculations are organized for us with the `anova()` *function in R, although the columns are in different places. For example:*

```
X1 <- rnorm(100)
X2 <- rnorm(100)
X3 <- rnorm(100)
Y <- 1 + 2*X1 + 3*X2 + 4*X3 + rnorm(100,0,4)

M1 <- lm(Y ~ X1 + X2 + X3)
round(summary(M1)$coef,2)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.56       0.37    1.49     0.14
## X1              1.81       0.35    5.16     0.00
## X2              2.52       0.37    6.90     0.00
## X3              3.89       0.36   10.86     0.00

M0 <- lm(Y ~ X1)
round(summary(M0)$coef,2)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.7       0.64    1.11     0.27
## X1               2.0       0.59    3.38     0.00
```

```
anova(M0, M1)
## Analysis of Variance Table
## Model 1: Y ~ X1
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     98 3948.7
## 2     96 1336.4  2    2612.2 93.821 < 2.2e-16 ***
```

In the output from `anova()`*, M0 is labelled "Model 1" and M1 is labeled "Model 2".*

- *The* `Res.Df` *column gives* $rdf_{M0} = 98$ *and* $rdf_{M1} = 96$*;*
- *The* `RSS` *column gives* $RSS_{M0} = 3948.7$ *and* $RSS_{M1} = 1336.4$*;*
- *The* `Df` *column gives* $rdf_{M1} - rdf_{M0} = 98 - 96 = 2$*;*
- *The* `Sum of Sq` *column gives* $RSS_{M0} - RSS_{M1} = 3948.7 - 1336.4 = 2612.2$*;*
- *And the* `F` *column gives*

$$F = \frac{(RSS_{M0} - RSS_{M1})/(rdf_{M1} - rdf_{M0})}{RSS_{M1}/(rdf_{M1})} = \frac{2612.2/2}{1336.4/96} = 93.821$$

- *Finally,* `Pr(>F)` *gives the probability that, if* H_0 *were true, we would see an F statistic bigger than 93.831. Since this p-value is very small (less than* 2.2×10^{-16}*), we reject* H_0*:* $M0$ *holds so M1 is the better model.*

DIAGNOSTICS

I am still a little bit confused regarding the H1 hat matrix.

Sometimes it is ambiguous to judge whether the results of diagnosis plots satisfy the assumptions.

I do not really understand how could I judge which model is better or the best. There are many information we can look to a model, such as "R-squared", "Residual plots", "AIC" and so on, I am not sure how to balance between these staffs.

I still have a hard time interpreting the diagnostic plots, specifically the Scale-Location plot. In the case where there's functional dependence of variance on y, what additional information would Scale-Location plot give compared to Residual vs Fitted plot?

In fact, except for the Normal Q-Q plot, I still don't know exactly how to judge the fit of the other diagnostic plots, although I have learned about it on the Internet, I don't know if my knowledge is correct.

Something that confuses me is whether the Scale-Location plot and Residuals vs Fitted plot shows whether the variance of the $\hat{e}$, ϵ, y or $\hat{y}$ is non-constant or constant. Also, I'm not sure how to derive variances and covariances of $\hat{e}$, ϵ, y or $\hat{y}$.

I am also still a little bit confused regarding the process for model selection when looking purely at diagnostic plots. I feel like I know what they mean much better now, as I previously mentioned, but I think it is extremely difficult sometimes to be able to tell which model actually looks better, especially when they are relatively similar in strength of fit.

How to analyze the diagnostic plots, especially the leverage examples (don't know how to tell high/low leverage and residual).

How do high-leverage points affect a linear model?

The H_1 matrix is just the hat matrix for the intercept-only model $y = \beta_0 + \epsilon$. This is the smallest model we would ever compare any other model to. When the intercept-only model holds, it means none of the other predictors are useful for predicting y.

Judging whether a particular fitted model is "good enough" is genuinely hard, there's nothing automatic about it. Part of the reason for this is that there are many ways to measure "good enough" and they don't all agree. In addition, how good is "good enough" (and which measures of "good enough" you should use) will depend on what substantive problem you are trying to solve by fitting a model.

For now we have been concentrating on the casewise diagnostic plots (the four plots that R gives us) because they are useful for assessing <u>whether the regression assumptions hold for the fitted model.</u> Those assumptions are

- ***Linearity:** is y a linear function of the columns of the X matrix, or do we need transformations, interactions or additional predictors?*
  - *We check this primarily by looking for functional patterns in the residual vs fitted plot. Does it look like the residuals are some sort of noisy function of $\hat{y}$, or is there no real pattern in the relationship between $\hat{e}$ and $\hat{y}$?*
- ***Constant Variance:** Does the variance of the residuals depend on y or $\hat{y}$?*
  - *We can check for this using the residual vs fitted plot (is the width of the cloud of residuals fairly constant as we go from left to right?) or the scale-location plot (is the "trend" in this plot horizontal, or does it increase or decrease in some way?*
- ***Normality:** Are the standardized residuals approximately N(0,1)?*
  - *We look primarily at the Normal QQ plot for this*

There is nothing about the model assumptions that suggest large leverage is bad but, as a practical matter, points with large leverage have a disproportionately large influence on the predictions $\hat{y}$, and so we would like to identify those points, since the fitted model will be overly sensitive to them. If they are already near the regression line (or regression surface) then we don't worry about them too much. If they are far from the regression line (or surface) then removing them may greatly affect (for better or for worse) the actual values of the $\hat{y}$'s. As explained in lecture 2 (and visible in the residuals vs. leverage plot):

- *Any leverage above $2\frac{p+1}{n}$ (i.e. larger than twice the average leverage) is considered 'large'*
- *Any data point with Cook's distance greater than about 0.5 is considered to have both large leverage and large residual, and hence to be influential on the values of $\hat{y}$.*

That is why we identify large leverage values or large Cook's distance values and bring the corresponding data points to our collaborator or client to discuss (are they real, and so we just

have to live with them? Or are they due to a correctable coding error or data collection error? Etc.)

We will discuss other measures of fit (R squared, AIC, etc.) when we discuss variable selection.

TRANSFORMATIONS

It seems like the transformations we've seen so far address skewness in the data. What transformations should we use to address kurtosis? (Are non-monotonic transformations a bad idea in this process?)

I'm having some difficulty understanding Variance-Stabilizing Transformations (covered in Lecture 4).

I still feel shaky on the theory behind VST, specifically determining when and where we should be using certain transformations. This feeling does extend towards the application of all these methods we are learning in the wild on data not from homeworks. I don't feel confident with providing justifications and understanding what assumptions we can make (and should make) when making linear models.

The use of transformations which can make the variance of residuals constant.

I do not really understand is the Box-Cox transformation. In addition to the derivation being difficult, I do not fully understand when we would want to apply this in a real world setting. In other words, what situation would we need Box-Cox instead of just doing a log or square root transform?

The concept of Box-Cox is not so intuitive and I am trying to understand how it really works. It was the first I learned the concept, so I am still trying to understand the fundamentals behind the concept and the graphical interpretation behind it.

I think I am bit confused about Box-Cox transformation and standardized residual.

Something that still mystifies me in this course so far is the concept of box-cox method. I believe as it is covered more and I see it come up somewhere it will make more sense, but I'm a bit confused on it.

I am really confused on how the box cox likelihood is generated in such a complicated formula.

How does the Box-Cox function choose the best lambda?

I don't really understand about the reducing leverage: powers of X, especially the Box-cox.

I actually have a confusion of the GIR variable in question 1. Based on the normal QQ plot and histogram of the GIR variable, it has a little long right tail and maybe a good or a little long left tail. Do I need to make a transformation for this variable and what kind of simple transformation can I use.

The simple transformation of variables can only deal with the situation that both left and right tails lay on the same side of the line? For instance, the Q-Q plot for variable GIR in question 1 does not look perfect. However, as I tried several simple transformation, such as sqrt, log, square, none of these transformation makes the plot looks better. So I think the transformation we would like to choose should based on the distribution of the data points. For instance, the Q-Q plot for PrizeMoney looks like a logarithm, so we choose log(PrizeMoney) instead to make it more normally distributed. Thus, if we need to improve the Q-Q plot for GIR, a more complicated transformation might works. Is it correct?

I often use a trial-and-error approach when fitting y and x1, x2... to determine the form of x (whether it is the combination of x1 and x2, or log(x1) or other forms), is there a more logical way to find its form?

How should we generally check which transformation we need to make? Seems like we need to check each column of data to see if they are normally distributed. What if the dataset is too large?

I could use more practice with knowing how and when to apply transformations because I feel like my previous stats courses glossed over the criteria for determining when a transformation is needed. I do understand the concept of applying transformations to achieve a more normal distribution of points, but I'm just not sure how to tell which transformations should be applied in situations like the PGA golf data where it's unclear sometimes. I think if a data set has a clear trend towards a universally identifiable pattern like a negative quadratic, it's easier to see that it should be transformed as such. In future problems if I encountered an amorphous blob/circle of data points in a scatterplot however, I wouldn't know how to proceed.

It is useful to divide transformations up into

- *Transformations to correct distributions*
- *Transformations to account for nonlinear relationships*

Only continuous variables should be considered for transformation. We would virtually never transform a dummy variable, a discrete variable, or a categorical variable, except to improve the interpretation of the coefficients. More on that later…

A transformation that you do for one of the above two purposes may help, or hurt, the other purpose. Getting a "good enough" and interpretable transformation is more art than science.

<u>Transformations to correct distributions</u>

These can be either transformations of X's or transformations of Y

- *Transformations of X's.*
  - *Usually the only thing we really care about is severe skewness in the empirical distribution of an X variable. Severe skewness means that some values of X in the data may lead to high leverage points, and we'd like to avoid that*
  - *So we would like a transformation that makes X more symmetric*
  - *The best thing to do is to eyeball the distribution and use the rules that*
    - *Right skew can often be fixed with a power of X between 0 and 1, or log(x)*
    - *Left skew can often be fixed with a power of X between 1 and ∞, or exp(x)*
    - *The Box-Cox method uses maximum likelihood to find the "optimal", but not interpretable, power of X:* **`boxCox(x ~ 1)`**
    - *It is generally better to go with powers of X that can be interpreted. E.g.:* $X^{-2}, X^{-1}, X^{-1/2}, \log(X), X^{1/2}, X, X^2, \exp(X)$
      This can often be done by guess-and-check using histograms or normal qq plots of X.
  - *We are generally less concerned with kurtosis*
    - *Leptokurtic distributions (both tails too short) generally aren't a concern*

        - *Platykurtic distributions (both tails too long) can lead to high-leverage points, but I do not know of simple transformations to fix them*
    - *Any (continuous) distribution can be transformed to near-perfect normality, as follows:*

      Let $\hat{F}(x)$ be the empirical CDF of X: $\hat{F}(x) = \frac{1}{n}\,\#\{X\ values \leq x\}$ and let $G(x)$ be the standard normal CDF: $G(x) = P[Z \leq x]$, where $Z \sim N(0,1)$. Then

      $$X_{new} = G^{-1}(\hat{F}(X))$$

      will have a nearly perfect normal distribution (it would be perfect if $\hat{F}(x)$ were replaced with the theoretical CDF $F(x)$). But this transformation is completely uninterpretable to a colleague, collaborator, or even another statistician, in terms of any features of the practical problem that led to the data in the first place.
- *Transformations of Y:*
    - *The only distribution-related reason to transform y is to improve the distribution of the <u>residuals</u> ϵ.*
    - *If the scale-location plot (essentially $\sqrt{Var(y)}$ vs $\hat{y}$) suggests nonconstant variance $Var(Y) = h(E[\hat{Y}])$, then one can try the variance-stabilizing transformation $g(y) \approx \int \frac{1}{\sqrt{h(u)}} du$ on y . In practice this can be difficult to calculate, so variance-stabilizing transformations aren't often used.*
    - *Eyeballing histograms or normal qq plots of standardized residuals $\hat{r}_i$ and trying guess-and-check with familiar pawers of y like $y^{-2}, y^{-1}, y^{-1/2}, \log(y), y^{1/2}, y, y^2, \exp(y)$ is simple and often works.*
    - *If this isn't working, feeding Box-Cox the fitted regression model can result in a power of Y that improves the distribution of ϵ (though it may not be interpretable):* **`boxCox(lm.fit)`**

<u>*Transformations to account for nonlinear relationships*</u>

These can also be either transformations of Y or X.

- *Transformations of X's.*
    - *Sometimes we want to transform X to get the functional relationship of X and Y right. This usually helps with distribution of ϵ as well. We can*
        - *try guessing a "good enough" transformation, by looking at a scatter plot of Y vs X, or*
        - *we can use nonparametric regression to get the transformation. The nonparametric regression transformation will be "better" mathematically, but harder to explain to anyone.*
    - *Added variable plots and marginal model plots can help guess the right transformation*
- *Transformations of Y.*
    - *Sometimes a transformation of Y can help to get the relationship between Y and X right.*

- *Oftentimes transforming X's, especially to get the relationship between Y and X's right, will also help with the distribution of ϵ.*
- *Guess and check with familiar powers, looking at a scatter plot of y vs X, can work.*
- *Inverse response plots can help guess the right power transformation.*

WHAT TO INCLUDE IN THE MODEL

I'm not very sure when we should include interaction terms in a model, and what the criteria of a good interaction terms would be.

Something that I am still a little confused about is the application of interaction terms, especially with multiple quantitative variables. I was confused about identifying good uses for them without negatively impacting the model. I remember covering in class that omitted variable bias has a far greater negative effect than the penalty for additional variables, but other than testing all interactions (while still following the hierarchy principle), if there is a quicker/better way other than substantive reasoning. I am used to conducting regression with interactions that involve a categorical or dummy variable, so the use, selection, and interpretation with quantitative interaction variables is something I hope to clarify over the semester.

Interactions are another kind of nonlinear transformation of X's but they involve more than one X at a time (they are products of X's).

Usually when you include an interaction term (or a power of X) in the model, you want to include all the related lower-order interactions and main effects (or lower powers of X). This "hierarchy principle" helps with the flexibility and interpretability of the model. The "" notation in R model formulas helps enforce/encourage the hierarchy principle.*

Two-way interactions X_iX_j can involve two dummy variables, one dummy and one continuous variable, or two continuous variables. If X_i is categorical with more than two categories, then the interaction is the product of all the dummies for the values of X_i with whatever X_j is. Similarly for three-way interactions $X_iX_jX_k$, four-way, etc.

In X_iX_j, when one of the X's is discrete, assessing whether the interaction is needed is part of Analysis of Covariance (ANCOVA). If both are discrete, assessing whether the interaction is needed is part of Analysis of Variance (ANOVA; confusingly the same name as for the table that organizes calculation of an F statistic). When both are continuous, the interaction is one of the terms in the formula for a two-dimensional quasdratic surface (e.g. a paraboloid or hyperboloid).

In most data that I have seen, three way interactions are rare and four way interactions hardly ever occur. So in practice I often check all two-way interactions, then maybe all three way interactions, and maybe just to be safe I'll verify that none of the four-way interactions are significant. We can do a partial F test where M0 is the model without the interactions and M1 is the model with the interactions, to see if all the interactions together help with the fit. Since we usually think of adding interactions last in the model, checking the t-statistics is also not a crazy way forward (although we do have to beward of collinearity between the interactions).

Soon we will talk about variable selection methods, which give us other, potentially better tools than the F test and t tests.

And of course, we should never stray far from substantive reasoning.

NOTHING YET

Not really, but I feel like it will gets harder in the later courses though... I'm a statistic major so I've seen most of the stuff so far but I'm still worrying about what comes after.

Glad to hear it. I'm sure some things later will be more mystifying!