Stat 27850/30850: Group mini-project # 2

Data The data set for the first project is the Ames housing data set, from
`http://www.amstat.org/publications/jse/v19n3/decock/AmesHousing.xls`
Some documentation on this data set:

- `http://jse.amstat.org/v19n3/decock/DataDocumentation.txt` — all variables & levels are listed and explained
- `http://jse.amstat.org/v19n3/decock.pdf` — additional information

This data set records information on every home property sale in Ames, Iowa in the range 2006–2010, recording variables such as square footage, number of bedrooms/bathrooms, whether the home has central air conditioning, fence type, etc. There are $n = 2930$ data points. The original data set has 79 covariates (many of these are categorical, e.g., roof type), as well as one response (`SalePrice`, the price at which the home sold).

On Canvas, we provide a script named `getdata_ameshousing.R` to download and clean the data so that it's ready for R and organized in a simple format. The variables you will have after running the script are:

- Response variable Y = log(`SalePrice`). Due to a highly skewed distribution
- Covariate matrix X with 48 columns (e.g., `Lot Area`, `Year Built`, `GarageQual`, etc). The command `colnames(X)` will identify the features that are included. See the first documentation link for an explanation of these variables.

Many of the covariates in the original data set are removed or simplified, to reduce the complexity of the analysis. Specifically,

- All quantitative variables (e.g., `1st Flr SF`) are included in their original form.
- For factor variables that are ordinal (for example, `ExterQual` with levels Poor, Fair, Average/Typical, Good, Excellent), we have recoded these as numerical, with 1 always coding the worst quality/level and higher numbers indicating better quality/level.
- Factor variables that are categorical without an obvious ordering (for example, roofing material type) have been removed.
- If there are NA values that actually indicate that the feature is absent (e.g., `Garage Qual` is NA if there is no garage, or `Lot Frontage` is NA if there is no lot frontage on the property), this level is coded as 0. Otherwise, data points with any NA values (missing values) are removed.

In total, there are $n = 2923$ data points and $p = 48$ covariates after these steps.
7 removed

Guidelines for group & report Please see guidelines for mini-project 1.

Assignment Your assignment is to use the available data to estimate and test the association between the covariates (or a subset of the covariates) and the response, and to explore interesting issues and challenges around these types of questions.

Specifically, please address the following three questions.

1. **Simpson's paradox.** This well-known statistical paradox occurs when the association between some X and some Y appears positive within a larger population, but appears negative within any subpopulation (see `https://en.wikipedia.org/wiki/Simpson%27s_paradox` for examples).

   While it's best known for categorical covariates (like in the Wikipedia example), in the continuous case a related phenomenon is when the coefficient on X_j switches sign if you control for other covariates X_k.

   Here is an example: if we regress log(Sale price) on year built + # of bedrooms, the coefficient on # of bedrooms is positive (as expected—a house with more bedrooms is more valuable). But if we regress also on the square footage of the house, the coefficient on # of bedrooms is negative.

```
> summary(lm(Y~X[,c(5,18)]))
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -4.9624465  0.3753080  -13.22   <2e-16 ***
X[, c(5, 18)]Year Built     0.0084629  0.0001896   44.65   <2e-16 ***
X[, c(5, 18)]Bedroom AbvGr  0.1050444  0.0069319   15.15   <2e-16 ***
>
> summary(lm(Y~X[,c(5,11,12,18)]))
Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                         2.680e-01  2.734e-01   0.980    0.327
X[, c(5, 11, 12, 18)]Year Built     5.584e-03  1.398e-04  39.960  < 2e-16 ***
X[, c(5, 11, 12, 18)]1st Flr SF     6.296e-04  1.162e-05  54.161  < 2e-16 ***
X[, c(5, 11, 12, 18)]2nd Flr SF     4.436e-04  1.154e-05  38.445  < 2e-16 ***
X[, c(5, 11, 12, 18)]Bedroom AbvGr -4.701e-02  5.813e-03  -8.087 8.91e-16 ***
```

   We can explain this as follows: given the size of the house, splitting it into more bedrooms reduces value because the rooms are too small / this indicates overcrowding / some other negative factor. On the other hand, not knowing the size of the house, having more bedrooms indicates a likely larger house.

   For this first question, answer the following:

   - Find a completely different example of this paradox, within this same data set, of a coefficient that changes sign, and give an intuitive interpretation (like the one above) for what you see.
   - Explain how this type of paradox/phenomenon can create issues in terms of interpreting the output of a selective inference procedure, run on a real data set.

2. **Selective inference.** Next, we will carry out a selective inference procedure on this data set. (You do not need to "correct" for the fact that you already viewed the data to answer question 1—you can think of each question as a separate study.)

   1 First, you may want to do an initial check for outliers or other data cleaning type issues.

   2 Next, for selection, you should use a simple rule like marginal screening or forward stepwise. (You can use Lasso if you choose, but all your calculations should be done on your own—please do not use existing tools like the selective inference package in R—so you may prefer to use something like marginal screening for the sake of simplicity. However, be sure that the method you choose is intuitively meaningful. For example, the value of $|X_j^\top y|$ is sensitive to issues like the scale of X_j, a nonzero mean of y, etc.) Reduce the outcome of your screening event to a set of linear inequalities, of the form $A \cdot y \leq b$. (Your work in Pset3 should help with this step.)

   3 Finally, calculate a p-value for each selected coefficient (with the given sign, i.e., a one-sided test). To run this method, theoretically we will need to have a known value of σ^2 in order to run this procedure—in practice, we need to plug in an estimate.

To help with the necessary calculations for the inference step, here is code to compute the truncated normal distribution, given a vector v chosen as a result of the selection event $A \cdot y \leq b$. In other words, if the initial distribution is $y \sim N(\mu, \sigma^2 \mathbf{I})$ and, after observing $A \cdot y \leq b$, we pick a vector v to test (i.e., we want to test $v^\top \mu$), the selective distribution for $v^\top y$ is a truncated normal, $v^\top y \sim TN(v^\top \mu, \sigma^2; [c_{lower}, c_{upper}])$, where the code below will compute the endpoints c_{lower} & c_{upper}. (You will need to provide A, b, σ^2, v, and the variable `prp_v_y` which is $\mathcal{P}_v^\perp(y)$, the orthogonal projection.)

```
c_lower = -Inf; c_upper = Inf
for(l in 1:dim(A)[1]){
# The l-th inequality constraint is: (A*y)_l <= b_l
# Equivalently: (A*prp_v_y)_l + (A*v)_l/sum(v^2) * (v'*y) <= b[l]
coef1 = sum(A[l,]*v)/sum(v^2)
coef2 = b[l] - sum(A[l,]*prp_v_y)
if(coef1>0){c_upper = min(c_upper,coef2/coef1)} # update the upper endpoint
if(coef1<0){c_lower = max(c_lower,coef2/coef1)} # update the lower endpoint
}
```

3. **Checking assumptions.** Finally, do you believe that this data set sufficiently satisfies the assumptions needed, for the selective inference method to give meaningful results? You can explore this from any angle you like. For example, you may consider issues like non-constant variance in the noise; clusters / different distributions within different subpopulations; non-independent data; etc. Your answer to this last question should involve an empirical exploration in addition to discussion. It would be very interesting to explore the types of issues that can be caused by the problems you identify—for example, if you suspect nonconstant variance, could you simulate what types of issues would arise with selective inference, in this case?