Lab: Linear Regression and Bias-Variance Tradeoff

1. Linear Regression - Bias-Variance Tradeoff

Assume that X_1, X_2 are two independent variables but with a same distribution $N(1,1)$. The true relationship between Y_i and X_{1i}, X_{2i} is $Y_i = 1 + X_{1i} + 0.001X_{2i} + \epsilon_i$, where $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$.

(a) Is the intercept/coefficient of X_1 biased if you only regress on X_1? What is your intuition?

The estimated intercept is biased but the coefficient of X_1 is not biased. Since X_1, X_2, ϵ are independent, we can treat $\eta = 0.0001X_2 + \epsilon$ as a new error term instead of ϵ. Based on the properties of normal distribution, we derive the distribution of η, which is $N(0.0001, 0.0001^2 + \sigma^2) = 0.0001 + N(0, 0.0001^2 + \sigma^2) = 0.0001 + \eta^*$.Thus, our model is equivalent to $Y = 1.0001 + X_1 + \eta^*$, where η^* has a mean 0. From here, we can see, if we only regress on X_1, the intercept would be biased but not the coefficient of X_1.

(b) Follow-up: Show your conclusion in (a) mathmatically. Here are some hints:

Step 1: Based on the model we fit, we assume $E(Y|X_1) = \beta_0 + \beta_1 X_1$.
Step 2: We know the 'true' relationship between Y and X_1, X_2, use it to replace Y in the above equation.

$$E(\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \epsilon | X_1) = \beta_0 + \beta_1 X_1$$

Step 3: Work on the expectation and reach your conculsion.

$$\begin{aligned} \alpha_0 + \alpha_1 E(X_1|X_1) + \alpha_2 E(X_2|X_1) + E(\epsilon|X_1) &= \beta_0 + \beta_1 X_1 \\ \alpha_0 + \alpha_1 X_1 + \alpha_2 E(X_2) + E(\epsilon) &= \beta_0 + \beta_1 X_1 \\ \alpha_0 + \alpha_1 X_1 + \alpha_2 &= \beta_0 + \beta_1 X_1 \\ (\alpha_0 + \alpha_2) + \alpha_1 X_1 &= \beta_0 + \beta_1 X_1 \end{aligned}$$

This equation is valid no matter what X_1 is, so $\beta_0 = \alpha_0 + \alpha_2$ and $\beta_1 = \alpha_1$.

(c) Would you include X_2 to improve your model based on your intuition?
No, the effect size of X_2 is too small. By estimating a new parameter, we may lose power in estimating the existing parameters (larger variance).

(d) [Teamwork] Verify your conclusion using a simulation. Please follow the comments in the code chunk.

```
## Step 1: generate a training set
set.seed(263)
n = 50
x1 = rnorm(n,1,1)
x2 = rnorm(n,1,1)
eps = rnorm(n,0,0.2)
y = 1+x1+0.001*x2+eps
trainset=data.frame(cbind(y,x1,x2))
## Step 2: fit models on trainset: y~x1 and y~x1+x2
fit1 = lm(y~x1,data=trainset)
fit2 = lm(y~x1+x2,data=trainset)
## Step 3: generate a test set
m=10000
x1 = rnorm(m,1,1)
```

```
x2 = rnorm(m,1,1)
eps = rnorm(m,0,0.2)
y = 1+x1+0.001*x2+eps
testset=data.frame(cbind(y,x1,x2))
## Step 4: get the predictions in test set:
pred1 = predict(fit1,testset)
pred2 = predict(fit2,testset)
## Step 5: compare the MSEs in test set
MSE1 = mean((testset$y-pred1)^2)
MSE2 = mean((testset$y-pred2)^2)
```

Try to answer the questions below and get the idea of bias-variance tradeoff:
(1) In Model 1, the estimate of the intercept is unbiased, the MSE on the test set is 0.0398.
(2) In Model 2, the estimate of the intercept is biased, the MSE on the test set is 0.0403.
(3) Based on MSE, Model 1 is better, so you can infer that the predicitons using Model 2 have a larger variance.

(e) (Optional advanced problem) Let's go back to (a) and think, is the intercept/coefficient of X_1 biased if you only regress on X_1, given that X_1 is correlated with X_2?

Both intercept and coefficient would be biased. The only different part in derivation from (b) is $E(X_2|X_1)$ is not 1 anymore, it's a function of X_1 now.

2. Predict House Price Using Regression

This dataset('kc_house_data.csv') contains house sale prices for King County, which includes Seattle. It includes homes sold between May 2014 and May 2015. There are 19 house features plus the price and the id columns, along with 21613 observations. The dictionary of the variables is listed in the next page.

```
## Load in the data, split the data into training set and test set
house = read.csv('kc_house_data.csv',header = T)
trainset = house[1:floor(nrow(house)*0.8),]
testset = house[-(1:floor(nrow(house)*0.8)),]
```

(a) Fit a linear model on the training set: price = bedrooms + bathrooms + condition. Interpret the estimated coefficient of bathrooms and provide the corresponding 95% confidence interval.

```
fit1 = lm(price~bedrooms+bathrooms+condition,data=trainset)
summary(fit1)
```

```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + condition, data = trainset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1482279  -180134   -39247   110776  5925764
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -198419      15333 -12.941  < 2e-16 ***
```

```
## bedrooms       12184       2956   4.122 3.77e-05 ***
## bathrooms     248936       3662  67.975  < 2e-16 ***
## condition      51341       3567  14.394  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 309800 on 17286 degrees of freedom
## Multiple R-squared:  0.2836, Adjusted R-squared:  0.2835
## F-statistic:  2281 on 3 and 17286 DF,  p-value: < 2.2e-16
```

```
confint(fit1)
```

```
##                   2.5 %     97.5 %
## (Intercept) -228473.190 -168364.75
## bedrooms       6390.309   17976.88
## bathrooms    241758.221  256114.63
## condition     44349.454   58332.40
```

Interpretation: The average house price will increase 248936 dollars for every one more bathroom in the house adjusting for number of bedrooms and house condition.

(b) Fit a linear model on the training set: price = bedrooms + bathrooms + condition + sqft_above. Compare the coefficients of bedrooms here with the one in (a), what do you find?

```
fit2 = lm(price~bedrooms+bathrooms+condition+sqft_above,data=trainset)
summary(fit2)
```

```
##
## Call:
## lm(formula = price ~ bedrooms + bathrooms + condition + sqft_above,
##     data = trainset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1047706  -161963   -32861   106841  5007967
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.894e+05  1.402e+04 -20.638  < 2e-16 ***
## bedrooms    -1.640e+04  2.728e+03  -6.011 1.89e-09 ***
## bathrooms    1.058e+05  4.089e+03  25.876  < 2e-16 ***
## condition    7.717e+04  3.271e+03  23.594  < 2e-16 ***
## sqft_above   2.246e+02  3.726e+00  60.265  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 281600 on 17285 degrees of freedom
## Multiple R-squared:  0.408,  Adjusted R-squared:  0.4078
## F-statistic:  2978 on 4 and 17285 DF,  p-value: < 2.2e-16
```

The coefficient of bedrooms becomes negative, which doesn't make sense. It is caused by the colinearity of bedrooms and sqft_above.

Columns

- # id a notation for a house
- A date Date house was sold
- # price Price is prediction target
- # bedrooms Number of Bedrooms/House
- # bathrooms Number of bathrooms/House
- # sqft_living square footage of the home
- # sqft_lot square footage of the lot
- # floors Total floors (levels) in house
- A waterfront House which has a view to a waterfront
- A view Has been viewed
- A condition How good the condition is (Overall)
- A grade overall grade given to the housing unit, based on King County grading system
- # sqft_above square footage of house apart from basement
- # sqft_basement square footage of the basement
- # yr_built Built Year
- # yr_renovated Year when house was renovated
- # zipcode zip
- # lat Latitude coordinate
- # long Longitude coordinate
- # sqft_living15 Living room area in 2015(implies-- some renovations) This might or might not have affected the lotsize area
- # sqft_lot15 lotSize area in 2015(implies-- some renovations)

Figure 1: variable_dict

(c) Using the model in (a), predict the price in the test set and calculate the mean square loss $L = \frac{1}{M}\sum_{i=1}^{M}(Y_i - \hat{Y}_i)^2$.

```
pred = predict(fit1,testset)
mean((pred-testset$price)^2)
```

```
## [1] 96659063399
```

(d) [Teamwork] Competition Time!
Use linear regression to get the best prediction! (Minimal square loss in test set). Think about:
1.Transformation: what is the best $\phi(x)$, e.g. log(sqft_above)? sqrt(sqft_above)? or original?
2.Should we include all variables? How to combine different pieces of information, e.g.yr_built and yr_renovated?
3.continuous or categorical?

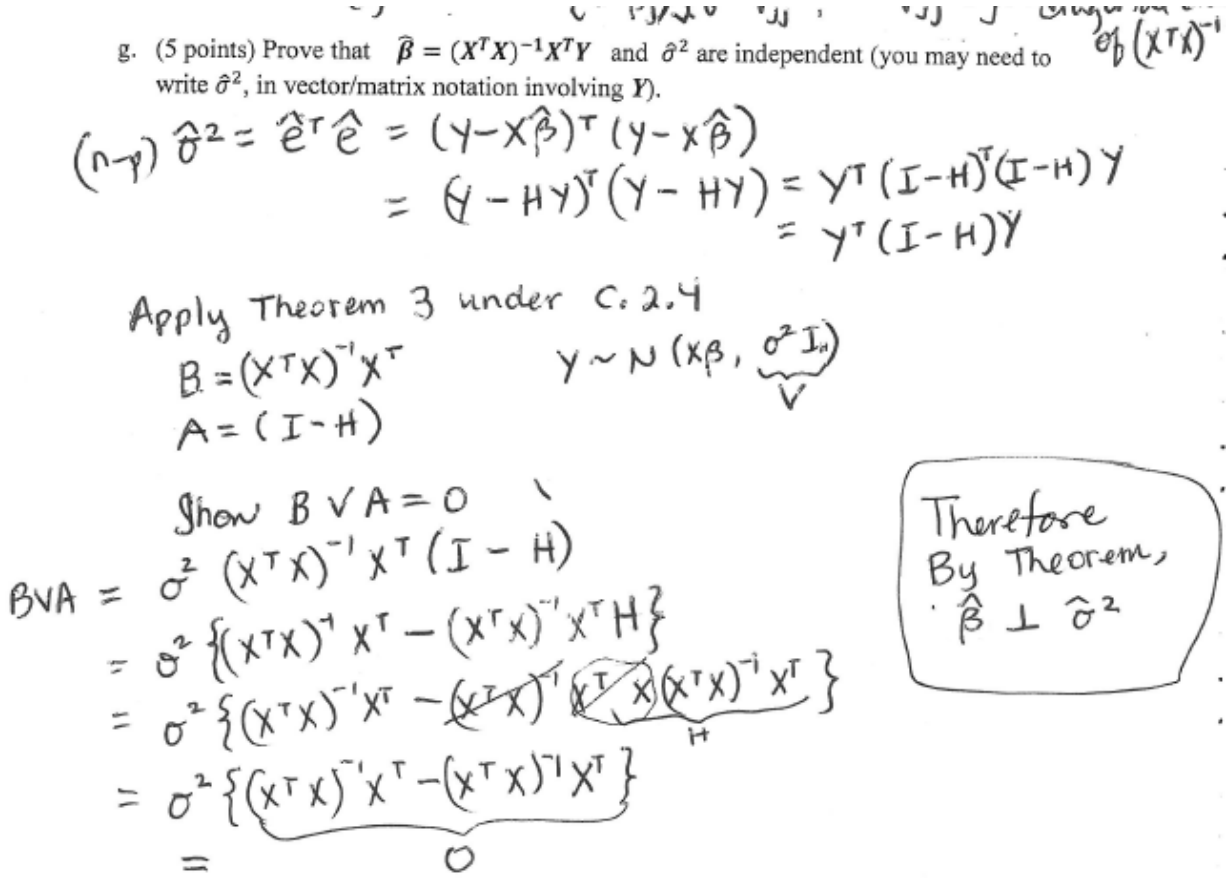

Figure 2: sol

3. (Optional advanced problem) Distribution Theory - Matrix Representation

This question is beyond the scope of this class. It is here only for those who want more practice on matrix representations.

Let y be a $k \times 1$ multivariate normal random vector with mean μ and nonsingular variance-convariance matrix V, $y \sim N(\mu, V)$. Additionally, let A be a $k \times k$ matrix of constants and B be a $q \times k$ matrix. Then, the linear form $W = By$ and quadratic form $U = y^TAy$ are independent if $BVA = 0$.

Prove that $\hat{\beta} = (X^TX)^{-1}X^TY$ and $\hat{\sigma}^2$ are independent using the theorem above.