HW3

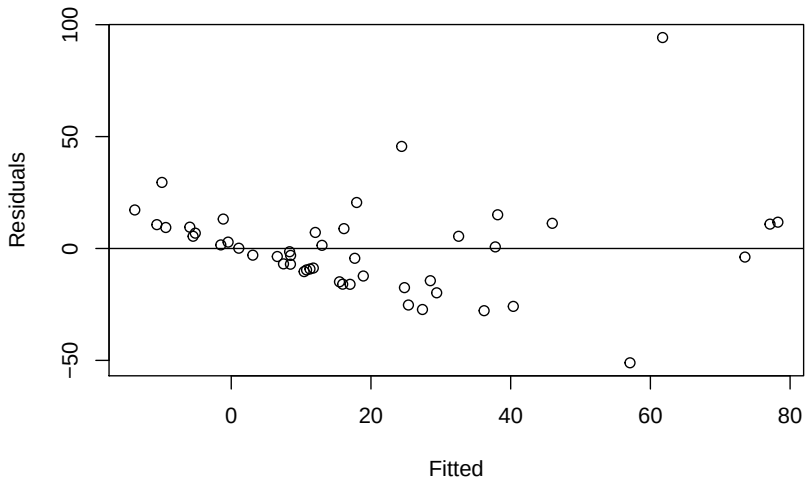
Jinhong Du - 12243476

2019/10/08

Contents

1. Using the `teengamb` dataset, fit a model with gamble as the response and the other variables as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevent. Do not provide any plots about which you have nothing to say.

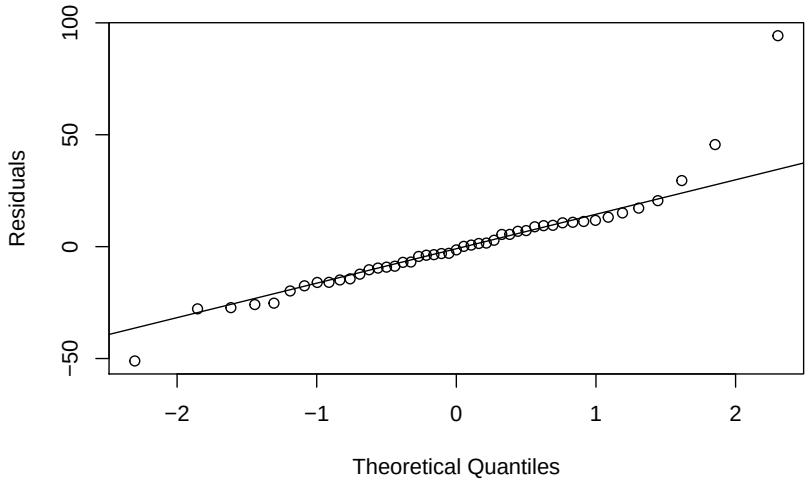
(a) Check the constant variance assumption for the errors.



The plot shows nonconstant variance and maybe a nonlinear (U-shape) relatinship exists.

```
library(faraway)
data(teengamb)
model <- lm(gamble ~ ., teengamb)
plot(fitted(model),residuals(model),xlab="Fitted",ylab="Residuals")
abline(h=0)
```

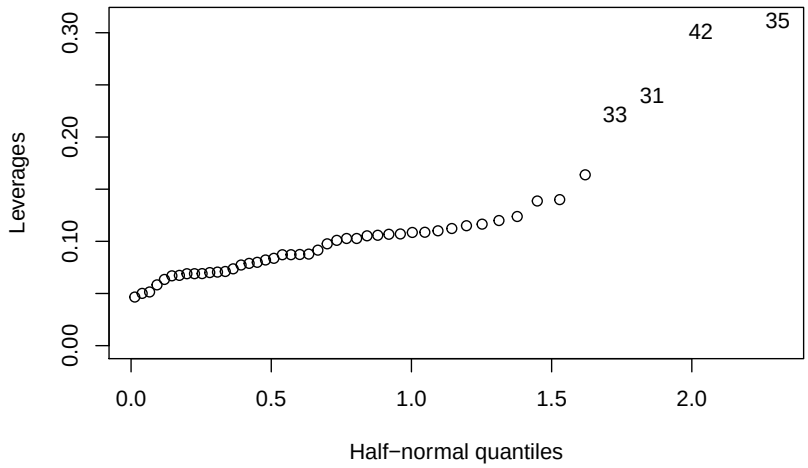
(b) Check the normality assumption.



The distribution of the residuals has heavier tail then normal distribution. So the normality assumption may be unsatisfied. Otherwise, the few points away from the QQ-line may be outliers.

```
qqnorm(residuals(model),ylab="Residuals",main="")
qqline(residuals(model))
```

(c) Check for large leverage points.



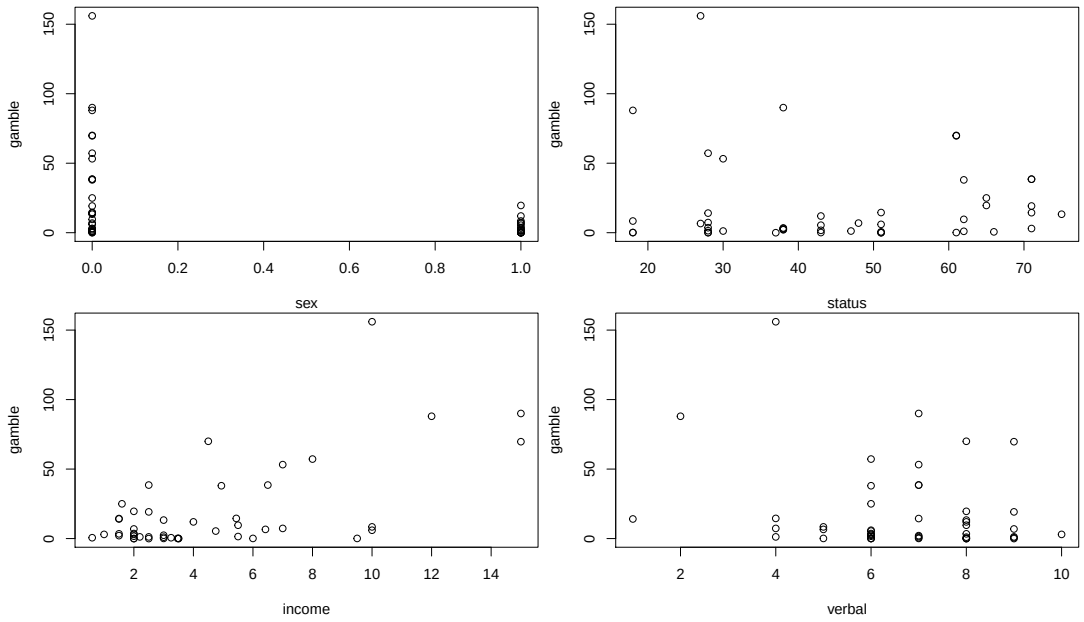
The leverage for the ith sample is h_{ii}. The half-norm plot below suggest that the 35th, 42th, 31th and 33th samples are large leverage points ($> \frac{2p}{n}$).

```
hatv <- hatvalues(model)
head(hatv)
```

```
##          1          2          3          4          5          6
## 0.07988226 0.10851291 0.06347643 0.10273955 0.13866946 0.16378563
```

```
n <- sum(hatv>length(model$coefficients)*2/dim(teengamb)[1])
r_name <- row.names(teengamb)
halfnorm(hatv,n,labs=r_name,ylab="Leverages")
```

(d) Check for outliers.



From the plots of the response versus each predictor, we can see there is one sample has extremely large value of the response. So it may be an outlier.

Four samples on the top right corner in the plot of gamble versus income may also be outliers.

Three samples on the top left corner in the plot of gamble versus verbal may also be outliers.

However, if we check for studentized residuals with Bonferroni correction, then the 24th sample is an outlier.

```
plot(gamble ~ sex, teengamb)
plot(gamble ~ status, teengamb)
plot(gamble ~ income, teengamb)
plot(gamble ~ verbal, teengamb)
stud <- rstudent(model)
which(as.vector(stud) > abs(qt(.05/(dim(teengamb)[1]*2), model$df.residual)))
```

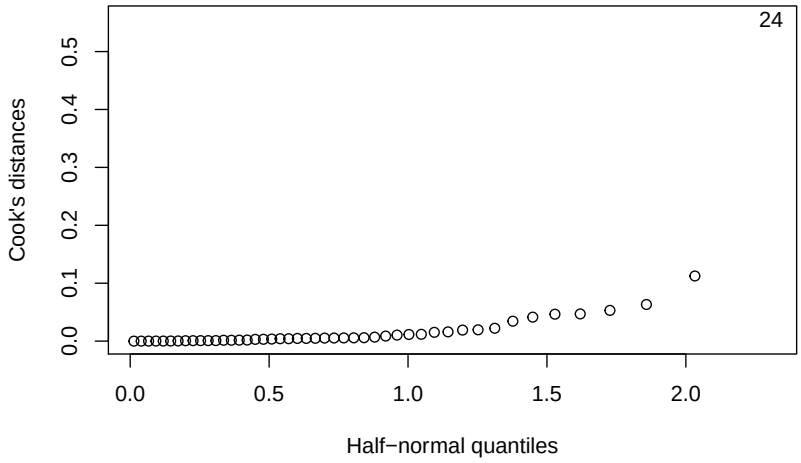
```
## [1] 24
```

(e) Check for influential points.

The Cook distances are defined as

$$D_i = \frac{(\hat{Y} - \hat{Y}_{(i)})^\top (\hat{Y} - \hat{Y}_{(i)})}{p\hat{\sigma}^2}$$

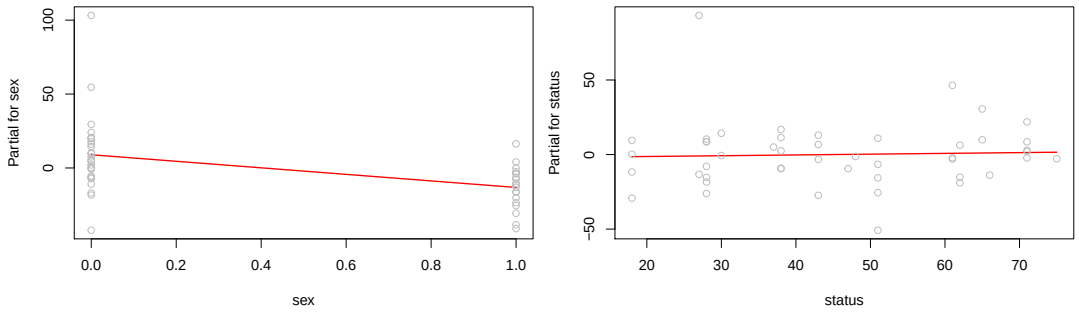
where $\hat{Y}_{(i)}$ is the predicted response for a fit without case i.

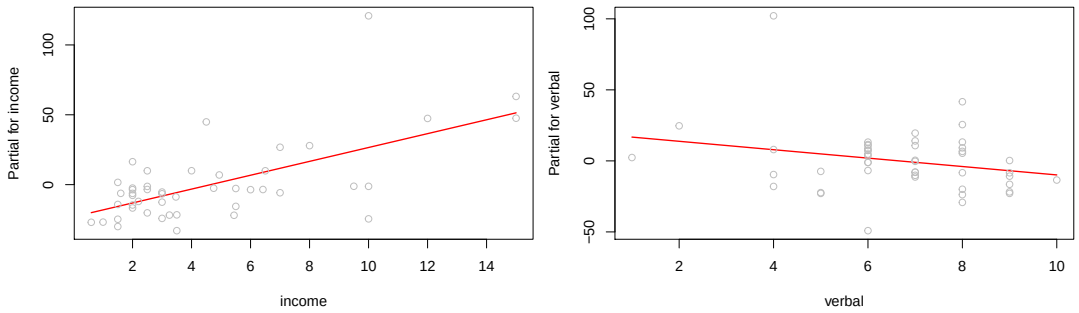


As we can see, the 24th sample is the influential point.

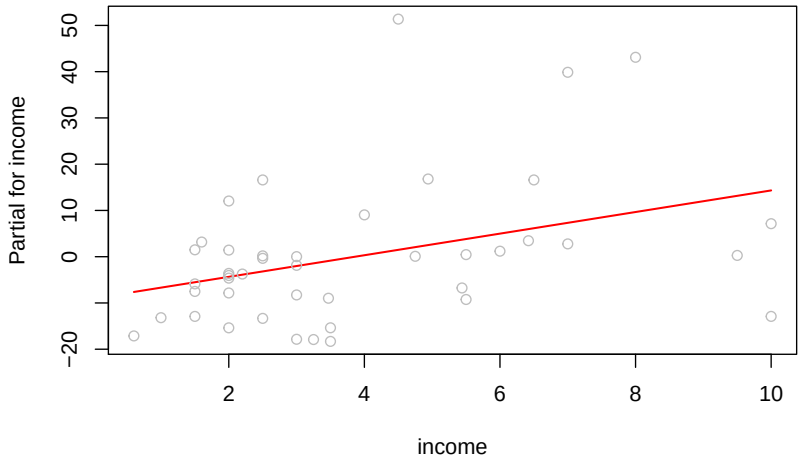
```
cook <- cooks.distance(model)
halfnorm(cook,1,labs=r_name,ylab="Cook's distances")
```

(f) Check the structure of the relationship between the predictors and the response.





From the plot of partial residuals of each income, we see that there might be some structure of the relationship. Maybe sex or high and low income will have different effect to the model. Let's remove the four potential outliers and refit again,



This new plot seems normal now. So there is no structure of relationship with income.

```
termplot(model, partial.resid=TRUE, terms=1)
termplot(model, partial.resid=TRUE, terms=2)
termplot(model, partial.resid=TRUE, terms=3)
termplot(model, partial.resid=TRUE, terms=4)

model_low_income <- lm(gamble ~ ., teengamb, subset = (income<11)&(gamble<100))
termplot(model_low_income, partial.resid=TRUE, terms=3)
```

2. We will work with the `sat` data set from Faraway. Each observation is one state, with $n = 50$ total. We will consider regressing `total` (average total SAT score) on `expend` (expenditure, i.e. public school funding per student), `ratio` (student-to-teacher ratio in public schools), `salary` (teacher salary), and `takers` (what proportion of eligible students take the SAT). (We will also include an intercept in every model.)

(a) Compare the coefficient on the `expend` covariate in the full model, against the coefficient if you regress `total` on `expend` only. Discuss what you see and explain intuitively what is happening in terms of the meaning of the variables.

The coefficient on the `expend` covariate in the full model is 4.4625942. The coefficient on the `expend` covariate in the reduced model is -20.8921737. The `expend` covariate seems to have opposite effect in these two models. This may be the reason that average SAT score is not mainly determined by `expend`. Other covariates like `takers` may be more important factor to influence average SAT score. It is misleading because of the omitted variables.

```
data(sat)
model <- lm(total ~ expend+ratio+salary+takers, sat)
cat('The coefficient of expend in full model is ', model$coefficients[1])
```

```
## The coefficient of expend in full model is  4.462594
```

```
reduced_model <- lm(total ~ expend, sat)
cat('The coefficient of expend in reduced model is ', reduced_model$coefficients[1])
```

```
## The coefficient of expend in reduced model is  -20.8921737
```

(b) Perform an F test, to test the full model against the model that regresses `total` on `takers` only. Do not use any R functions aside from `lm` and summary(lm(...)), and show all your calculations in R.

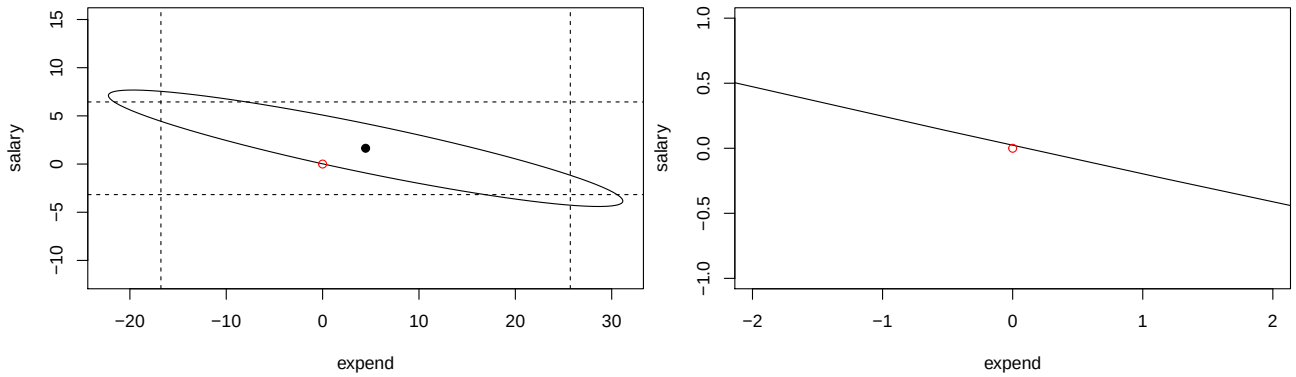
$$F = \frac{\frac{RSS_{reduced} - RSS_{full}}{df_{reduced} - df_{full}}}{\frac{RSS_{full}}{df_{full}}} \sim F_{df_{reduced}-df_{full}, df_{full}}$$

The F statistic is 3.213347 , which is larger than 2.811544 , the 95% quantile of F(3 , 45). The p-value is $0.03164874 < 0.05$, so we should reject the null hypothesis that $\beta_{ratio} = \beta_{salary} = \beta_{takers} = 0$.

```
reduced_model <- lm(total ~ takers, sat)
df_full <- summary(model)$df[2]
df_reduced <- summary(reduced_model)$df[2]
RSS_full <- summary(model)$sigma^2 * df_full
RSS_reduced <- summary(reduced_model)$sigma^2 * df_reduced
F_statistics <- (RSS_reduced-RSS_full)/(df_reduced-df_full)/(RSS_full/df_full)
cat('The F statistic is ',F_statistics,', which is larger than ',
    qf(0.95,df_reduced-df_full,df_full),', the 95% quantile of F(',
    df_reduced-df_full,',',df_full, ').\n',
    'The p-value is ',1-pf(F_statistics,df_reduced-df_full,df_full),
    '< 0.05.')
```

```
## The F statistic is  3.213347 , which is larger than  2.811544 , the 95% quantile of F( 3 , 45 ).
##  The p-value is  0.03164874 < 0.05.
```

(c) In the full model, draw the confidence region (i.e. the ellipse) for the coefficients ($\beta_{salary}, \beta_{expend}$). (You can use the `ellipse` library, see Faraway section 3.4 for an example—you can use the default confidence level of 95%). Explain the resulting ellipse shape that you see (hint: look at the correlations between the predictors).



Since the origin lies outside the ellipse region, we should reject the null hypothesis that $\beta_{salary} = \beta_{expend} = 0$. The shape of ellipse region shows that the estimated variance of β_{expend} is much larger than that of β_{salary}. Also, $\hat{\beta}_{expend}$ and $\hat{\beta}_{expend}$ seem to have slight negative linear relationship.

```
require(ellipse)
plot(ellipse(model,c(2,4), level=0.95),type="l", asp=1)
points(coef(model)[2], coef(model)[4], pch=19)
abline(v=confint(model)[2,],lty=2)
abline(h=confint(model)[4,],lty=2)
points(0, 0, pch=1, col='red', lwd=1)
```

3. In this problem we will do a simulation to examine how errors in our assumptions affect various calculations in regression.

(a) Generate data for a simple linear regression as follows: use sample size $n = 100$, generate the covariate values from a Uniform[0, 1] distribution and generate response values as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \text{ where } \epsilon_i \overset{iid}{\sim} N(0, 1)$$

with $\beta_0 = \beta_1 = 1$. Run this simulation 1000 times, each time storing (1) the estimate $\hat{\beta}_1$ and (2) the estimate of its (square root) variance, SE($\hat{\beta}_1$). (Hint: use `summary(model)$coef`, this matrix will have entries corresponding to $\hat{\beta}_1$ and its SE that you can easily extract for each run of your simulation). Compare (i) the empirical mean of $\hat{\beta}_1$ versus the target value β_1, and (ii) the empirical standard deviation of $\hat{\beta}_1$, versus the median value of SE($\hat{\beta}_1$). Is $\hat{\beta}_1$ unbiased? Does the estimated SE of $\hat{\beta}_1$ match the observed variation? Explain what you see.

The empirical mean of $\hat{\beta}$ is 1.003521 and is near to $\beta = 1$. The empirical standard deviation of $\hat{\beta}$ is 0.3473913 and is near to the median of $SE(\hat{\beta}) = 0.3464273$. $\hat{\beta}_1$ is unbiased, and the estimated SE of $\hat{\beta}_1$ matches the observed variation. Since the underlying model is exactly a Gaussian linear model, so theoretically $\hat{\beta}_1$ is unbiased and

$$SE(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \approx \sqrt{\frac{1}{100 \cdot Var(U)}} = \frac{1}{\sqrt{\frac{100}{12}}} \approx 0.3464102.$$

```
set.seed(0)
n_simulation <- 1000
result <- matrix(0, n_simulation, 2)

for (i in c(1:n_simulation)) {
    n <- 100
    X <- runif(n)
    epsilon <- rnorm(n)
    Y <- 1 + 1 * X + epsilon
    df <- data.frame(X=X, Y=Y)
    model <- lm(Y~X, df)
    result[i, ] <- summary(model)$coef[2, c(1,2)]
}
cat('The empirical mean of beta_hat is ', mean(result[,1]), '.\n',
    'The empirical standard deviation of beta_hat is ', sd(result[,1]), '.\n',
    'The median of standard deviation of beta_hat is ', median(result[,2]), '.')
```

```
## The empirical mean of beta_hat is  1.003521.
## The empirical standard deviation of beta_hat is  0.3459569.
## The median of standard deviation of beta_hat is  0.3473913.
```

(b) Now repeat with a different data generating mechanism,

$$Y_i = \beta_0 + \beta_1 X_i + (X_i)^4 \cdot \epsilon_i \text{ where } \epsilon_i \overset{iid}{\sim} N(0, 1)$$

which has a linear mean but heteroskedastic variance. Is $\hat{\beta}_1$ unbiased? Does the estimated SE of $\hat{\beta}_1$ match the observed variation? Explain what you see.

$\hat{\beta}_1$ is unbiased and the estimated SE of $\hat{\beta}_1$ does not match the observed variation. This is because that the term $X_i^4 \in (0, 1)$ which squeezes the error term, so that it reduce the estimated varaince.

```r
set.seed(0)
n_simulation <- 1000
result <- matrix(0, n_simulation, 2)

for (i in c(1:n_simulation)) {
    n <- 100
    X <- runif(n)
    epsilon <- rnorm(n)
    Y <- 1 + 1 * X + X^4 * epsilon
    df <- data.frame(X=X, Y=Y)
    model <- lm(Y~X, df)
    result[i, ] <- summary(model)$coef[2, c(1,2)]
}
cat('The empirical mean of beta_hat is ', mean(result[,1]), '.\n',
    'The empirical standard deviation of beta_hat is ', sd(result[,1]), '.\n',
    'The median of standard deviation of beta_hat is ', median(result[,2]), '.')
```

```
## The empirical mean of beta_hat is  1.001186 .
## The empirical standard deviation of beta_hat is  0.1647446 .
## The median of standard deviation of beta_hat is  0.113025 .
```

(c) Next, using the same heteroskedastic-variance model, run your simulation again. For each run, use the fitted model construct a prediction interval at $x = 0.1$ at level $1 - \alpha = 0.9$. Next generate a Y value at this X value, drawn from the same model, and record whether or not it lands inside the prediction interval. What proportion of your trials succeed, i.e. what proportion of the time does the prediction interval contain the new Y value? Then repeat at $x = 0.9$. What proportion of the time does the prediction interval contain the new Y value? Explain what you see.

For $X = 0.1$, the proportion of the time that the prediction interval contain the new Y value is 0.9 . For $X = 0.9$, the proportion of the time that the prediction interval contain the new Y value is 0.587. In this heteroskedastic-variance model, the effect of the variance is reduced by the term X_i^4. When $X = 0.1$, the effect is much smaller than the one when $X = 0.9$. So at $X = 0.1$, the generated data will be more concentrated near the line $Y = 1 + X$, and the coverage rate of the prediction interval is larger.

```r
set.seed(0)
n_simulation <- 1000


for (new_X in c(0.1,0.9)) {
    result <- matrix(0, n_simulation, 1)
    for (i in c(1:n_simulation)) {
        n <- 100
        X <- runif(n)
        epsilon <- rnorm(n)
        Y <- 1 + 1 * X + X**4 * epsilon
        df <- data.frame(X=X, Y=Y)
        model <- lm(Y~X, df)

        pred <- predict(model, data.frame(X=new_X), interval='prediction', level=0.9)
        y_true <- 1 + 1 * new_X + new_X^4 * rnorm(1)
        if((y_true>pred[2]) && (y_true<pred[3])){
            result[1] <- 1
        }
```

```
    }
    cat('For X =',new_X, ', the proportion of the time that the prediction interval',
        'contain the new Y value is ', mean(result), '.')
}
```

```
## For X = 0.1 , the proportion of the time that the prediction interval contain the new Y value is  1.
## For X = 0.9 , the proportion of the time that the prediction interval contain the new Y value is  0.587.
```

4. Suppose that we have a response $Y = (Y_1, \ldots, Y_n)$ and a single covariate $X = (X_1, \ldots, X_n)$. Our data set of size n is a mixture of data from two populations, labeled 0 and 1 arbitrarily, e.g. data from public schools and private schools. Let P_i be 0 or 1 to indicate which population the ith data point came from. Let n_0 and n_1 be the number of data points from each population, with $n_0 + n_1 = n$. If we're not sure whether the association between X & Y is the same within the two groups, we might do one of the following:

1. **Option 1: Split the data into two parts—one data set of size n_0 containing all the data points from population 0, and the other of size n_1 containing all data points from population 1. We could then run two linear regressions**

$$\textbf{Data from population 0: } Y_i = \beta_0^{(0)} + \beta_1^{(0)} X_i + \textbf{noise}$$

and

$$\textbf{Data from population 1: } Y_i = \beta_0^{(1)} + \beta_1^{(1)} X_i + \textbf{noise}.$$

2. **Option 2: Run a single linear regression to fit the model:**

$$\textbf{Combined data from both populations: } Y_i = \beta_0 + \beta_1 X_i + \beta_2 P_i + \beta_3 (X_i \cdot P_i) + \textbf{noise}.$$

(a) Are these two options the same or different in terms of what we're assuming about the mean of the response within this combined data set? Explain by comparing the coefficients $\beta_0^{(0)}, \beta_1^{(0)}, \beta_0^{(1)}, \beta_1^{(1)}$ from Option 1 with the coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ in Option 2.

Assuming that the mean of random noise is 0, then given X_i for data from population 0:

$$\mathbb{E}(Y_i | X_i) = \beta_0^{(0)} + \beta_1^{(0)} X_i;$$

for data from population 1:

$$\mathbb{E}(Y_i | X_i) = \beta_0^{(1)} + \beta_1^{(1)} X_i;$$

and for combined data from both population:

$$\begin{aligned} \mathbb{E}(Y_i | X_i) &= \beta_0 + \beta_1 X_i + \beta_2 P_i + \beta_3 (X_i \cdot P_i) \\ &= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_i & , \text{ if } P_i = 1 \\ \beta_0 + \beta_1 X_i & , \text{ if } P_i = 0 \end{cases} \end{aligned}$$

Then we have

$$\begin{cases} \beta_0^{(0)} = \beta_0 \\ \beta_1^{(0)} = \beta_1 \\ \beta_0^{(1)} = \beta_0 + \beta_2 \\ \beta_1^{(1)} = \beta_1 + \beta_3 \end{cases}$$

So these two options are the same. in terms of what we're assuming about the mean of the response within this combined data set.

(b) Are these two options the same or different in terms of what we're assuming about the variance of the response within this combined data set? Explain.

Option 1 allows different error variances for each population Option 2 assumes homogeneity of error variance across groups.

(i) If error terms in population 0 and 1 have the same variance, then model setting of option 2 is valid and the two options would have same estimate of σ. In this case, the dummy variable model yields a more precise estimate of the common variance than any of the separate regressions of option 1, since the former is based on more degrees of freedom.

(ii) If the error variances for two populations can not be assumed homogeneous, the model of option 1 is valid, but option 2 is not. The σ^2 obtained in option 2 will be in the middle of the estimates of error variance from two seperate models of option 1 and denpend on the relative size of the two populations.