QBS 121, Winter 2022: Assignment on Linear Models Part I

January 4, 2022

Do the circled questions

1 Problems

[✓ 1.] a. Show that the sample mean $\sum_{i=1}^{n} x_i/n$ minimizes the average squared distance, $\sum(x_i - \mu)^2$.

b. Show that the median minimizes the average distance, $\sum |x_i - \mu|$

2. Determine how translating and rescaling a predictor, X_i in a linear model, $Y = b_0 + b_1 X_1 + \cdots + b_k X_k + \epsilon$ effects the coefficient, b_i. That is, how do the coefficients change if one regresses on $X_i' = r + sX_i$ instead? If you find the math abstract, try it in practice (e.g. using lm()). List at least one motivations for rescaling a variable.

3. Let $\hat{\epsilon}_i = Y_i - (\hat{a} + \hat{b}X_i)$ be the residuals resulting from a least squares regression of Y on X. Sketch a proof that the sum of the residuals is zero and that $\sum \hat{\epsilon}_i x_i = 0$ (e.g. the independent variable and estimated residual are not "correlated").

[✓ 4.] Given a dependent variable Y and features $X_1, \ldots, X_k$ find the linear combination of the features that maximizes the correlation with Y.

[✓ 5.] How does R^2 change if (a) the dependent variable Y is rescaled, or (b) a new predictor $X_3 = aX_1 + bX_2$ is added to the linear model $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_2 + \epsilon$.

6. Suppose the number of covariates (independent variables) in your data set is equal to or more than the sample size minus one. How well can you fit the data? If you want to avoid the math try some examples in R with lm().

7. Suppose the true model describing the dependence of Y on X is $Y = \alpha_0 + \alpha_1 X + \alpha_2 X^2 + \epsilon$ where $\mathrm{E}X = \mu$ and the $\mathrm{Var}X = \sigma_X^2$ but you Y using the linear model $\mathrm{E}[Y|X] = \beta_0 + \beta_1 X$. Determine value of β_1 that optimizes the fit.

2 Data Analyses

2.1 Analysis of the npk dataset on agricultural yield

This is a dataset from agriculture for which alot of statistical methodology was developed 80 years ago. The data npk is part of base R. The dependent variable is *yield.* There are three exposures of interest, each binary, *N, P* and *K*, in addition to a variable called block.

1. Comment on the distribution of yield.
2. Regress yield on N, P and K, one at a time (univariable models).
3. Use a t-test to compare yield between observations in which N=1 (nitrogen present) and N=0). Compare this result to the regression of yield on N above.
4. Run a multivariable model of yield on the main effects, N, P and K. Interpret the coefficients.

2.2 Analysis of the stackloss dataset

This is data from a chemical production process. The dependent variable is *stack.loss*.

1. Comment on the distribution of stack.loss.
2. Regress yield on Air.Flow, Water.Temp and Acid.Conc., one at a time (univariable models).
3. Calculate a Pearson correlation of Air.Flow with stackloss and compare this result to the univariable regression of stackloss on Air.Flow above.
4. Run a multivariable model of stackloss on all three variables. Interpret the coefficients.
5. This part illustrates the explicit formula for least squares estimation.
   (a) Create the ”design matrix” or ”model matrix” corresponding to the main effects for Air.Flow, Water.Temp and Acid.Conc. e.g X ¡- cbind(1, Air.Flow,Water.Temp,Acid.Conc)
   (b) Calculate pred ¡- X %*% solve(t(X) %*% X) %*% t(X) %* % yield
   (c) Compare pred with the predicted values when you run lm(yield Air.Flow + Water.Temp + Acid.Conc)

3 Simulate and Analyze

Create a dataset that simulates a two-arm parallel randomized controlled trial with before and after intervention measurements of a continuous endpoint as follows:

1. Choose a sample size, say $n = 400$.
2. Generate a baseline version of the endpoint Y_0 that has a standard deviation of 10.
3. Let X be a binary variable indicating randomization to treatment or not (with 50-50 frequency).
4. Generate a post-intervention version of the endpont, Y_1 that has a correlation of $\rho = 0.7$ with Y_0. This can be done using $Y_1 = \rho Y_0 + \sqrt{1-\rho^2}N(0,\sigma) + \beta X$ where $\sigma = 10$ is the standard deviation of Y_0 and β is treatment effect. Try $\beta = 0.3$.
5. Verify that the Pearson correlation between the baseline and post-baseline versions is approximately 0.7.

Now analyze the data.

1. Test if the post-intervention endpoint (Y_1) is different between the two arms (i.e. the two levels of X).
2. Test if the change from baseline ($Y_1 - Y_0$) is different between the two arms.
3. Compare the change from baseline ($Y_1 - Y_0$) between the two arms *controlling* for the baseline value, Y_0.
4. Compare the post-intervention endpoint (Y_1) between the two arms *controlling* for the baseline value, Y_0.
5. Compare and commment on the last two adjusted analyses.

4 Simulations

1 a. Simulate two variables, X_1 and X_2, whose joint distribution is the bivariate normal with means of 1 and 2 respectively, standard deviations of 3 and 4, respectively and correlation of 0.5. Use a sample size of 500. **b**. Calculate the Pearson correlation of the two simulated variables. **c.** Calculate the R^2 when X_2 is regressed on X_1. **d.** Calculate the R^2 when X_1 is regressed on X_2. **e.** Comment on the values reported for parts b,c and d.

2. Run the code below and explain what it is trying to illustrate.

```
n <- 10
R <- 1000
R.sq <- adj.R.sq <- rep(NA, R)
for (r in 1:R) {
  X1 <- rnorm(n)
  X2 <- rnorm(n)
  X3 <- rnorm(n)
  X4 <- rnorm(n)
  Y <- rnorm(n)
  os <- summary(lm(Y ~ X1 + X2 + X3 + X4))
  R.sq[r] <- os$r.squared
  adj.R.sq[r] <- os$adj.r.squared
  }
mean(R.sq)
mean(adj.R.sq)
```

3. Run the code below and comment on the arguments to the Sim function and guess at what it is tryng to illustrate.

```
Sim <- function(n, beta, rfunc=rnorm, R=10^4) {
  N <- 2*n
  qt.975 <- qt(0.975, df=N-2)
  beta.est <- p.v <- rep(NA, R)
  CI <- matrix(nrow=R, ncol=2)
  for (r in 1:R) {
    X <- rep(0:1, each=n)
    noise <- rfunc(N)
    Y <- beta * X + noise
    os <- summary(lm(Y ~ X))
    beta.est[r] <- os$coef[2,1]
    CI[r,] <- os$coef[2,1] + os$coef[2,2] * qt.975 * c(-1,+1)
    p.v[r] <- os$coef["X", "Pr(>|t|)"]
    }
  mn <- mean(beta.est)
  cover <- mean(CI[,1]<beta & beta < CI[,2])
  emp.type1err <- mean(p.v < 0.05)
  par(mfrow=c(1,2))
  hist(beta.est)
  hist(p.v)
  c(average=mn, coverage=cover, emp.type1err=emp.type1err)
  }
Sim(n=20, beta=0, rfunc=rnorm)
Sim(n=20, beta=0, rfunc=rexp)
Sim(n=20, beta=0.5, rfunc=rnorm)
Sim(n=20, beta=0.5, rfunc=rexp)
Sim(n=50, beta=0, rfunc=rnorm)
Sim(n=50, beta=0, rfunc=rexp)
Sim(n=50, beta=0.5, rfunc=rnorm)
Sim(n=50, beta=0.5, rfunc=rexp)
```

4. The uniform distribution is symmetric but not bell-shaped. Generate 5 random variables, $X_1, X_2, \ldots, X_5$, that are uniformly distributed (U[0,1]) and independent. Created histograms of $X_1, X_1+X_2, \ldots, X_1+\cdots+X_5$. Comment on the shape as you add more of these random variables together.