HW2

Jinhong Du - 12243476

2019/10/08

Contents

1. Coefficients in simple linear regression vs multiple linear regression.
Suppose that there are two covariates, X_1 and X_2, which are generated from a bivariate normal distribution with correlation ρ. Assume that a normal linear model holds for Y, so that our observations $i = 1, \dots, n$ follow the distribution

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \textbf{ where } \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

(a) Run a simulation of this problem, and find choices of the parameters α, β_0, β_1, β_2, σ^2 such that:

- **If you fit a linear model of Y on covariate X_1 only, then the fitted slope is generally positive,**
- **But if you fit a linear model of Y on both covariates X_1 and X_2, then the coefficient β_1 on X_1 is generally negative.**

We select $\rho = 0.8$, $\beta = \begin{pmatrix} 1 \\ -1 \\ 10 \end{pmatrix}$ and $\sigma^2 = 1$. If fitting a linear model of Y on X_1 only, then $\hat{\beta}_1 > 0$ in all simulations. If fitting a linear model on X_1 and X_2, then $\hat{\beta}_1 < 0$.

```
library(MASS)
set.seed(0)
num_simulation <- 100
beta_1 <- c()
beta_2 <- c()
for (i in 1:num_simulation) {
    rho <- 0.8
    n <- 1000
    X <- mvrnorm(n, mu = c(0,0), Sigma = matrix(c(1,rho,rho,1), 2, 2) )
    X <- cbind(matrix(1, n, 1), X)

    beta <- matrix(c(1, -1, 10), 3, 1)
    sigma <- 1
    epsilon <- rnorm(n, 0, sigma^2)

    Y <- X %*% beta + epsilon
    df <- data.frame(Y=Y,X1=X[,2],X2=X[,3])
    model <- lm(Y~X1, df)
    model_2 <- lm(Y~X1+X2, df)
    beta_1[i] <- model$coefficients[2]
    beta_2[i] <- model_2$coefficients[2]
}
cat(mean(beta_1>0))
```

```
## 1
```

```
cat(mean(beta_2<0))
```

```
## 1
```

(b) Give a concrete example of three variables X_1, X_2, Y where you might plausibly expect to see this kind of trend, and explain. Your variables should be intuitive and common quantities, such as income, height, test score, etc.

In one family, let X_1 be the expense in any one month, X_2 be the income per person in the same month, and Y be the net income in the same month. Then $Y = n \times X_2 - X_1$ where n is the number of people in this family.

Obviously, X_1 and X_2 have positive correlation since more people tend to have more expense. Also, Y is mainly dominated by X_2 as X_1 are basically the same in different months. When fitting Y on X_1 only, X_2 is partially accounted by X_1, so $\hat{\beta_1}$ is positive. When fitting Y on X_1 and X_2, as the true relationship is linear, $\hat{\beta_1}$ is generally negative and near to the ture value -1.

2. For the prostate data, fit a model with lpsa as the response and the other variables as predictors.

(a) Compute 90 and 95% CIs for the parameter associated with age. Using just these intervals, what could we have deduced about the *p*-value for age in the regression summary?

Let $X_1, \dots, X_8$ denote `lcavol`, `lweight`, `age`, `Ibph`, `svi`, `lcp`, `gleason`, `pgg45`, respectively. The estimated coefficient associated with age is $\hat{\beta_3} = -0.0196372$. The 90% confidence interval of β_3 is (-0.0382102, -0.0010642). The 95% confidence interval of β_3 is (-0.0418406, 0.0025663). Using just these intervals, the p-value for age in the regression summary should be larger than 0.05 and smaller than 0.10, since $\beta_3 = 0$ in the null hypothesis lie in the 95% confidence interval but not in the 90% confidence interval.

```r
library(faraway)
data(prostate)
model <- lm(lpsa~., prostate)
summary(model)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

```r
confint(model, 'age', level=0.9)
```

```
##            5 %         95 %
## age -0.0382102 -0.001064151
```

```r
confint(model, 'age', level=0.95)
```

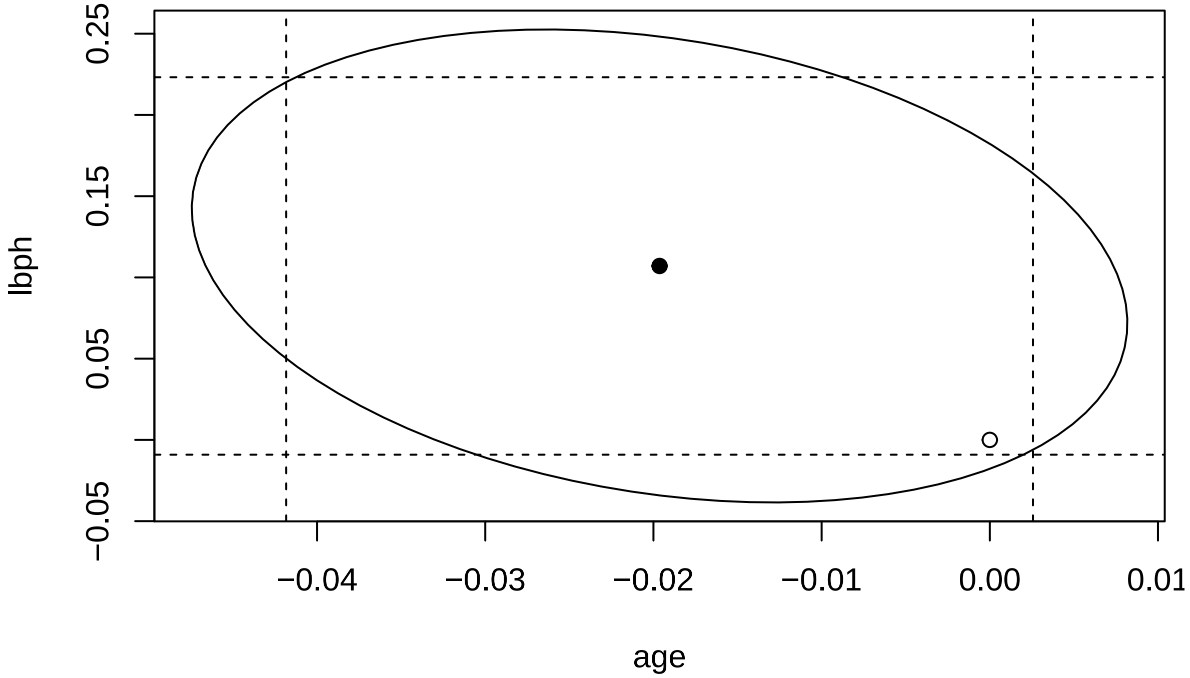
```
##           2.5 %      97.5 %
## age -0.04184062 0.002566267
```

(b) Compute and display a 95% joint confidence region for the parameters associated with age and Ibph. Plot the origin on this display. The location of the origin on the display tells us the outcome of a certain hypothesis test. State that test and its outcome.

The joint $100 \times (1-\alpha)\%$ confidence interval is given by

$$(\hat{\beta}-\beta)^\top X^\top X(\hat{\beta}-\beta) \leq p\hat{\sigma}^2 F_{p,n-p}^{(1-\alpha)},$$

which is a ellipsoidally shaped region in the following plot, where $F_{p,n-p}^{(1-\alpha)}$ is the $1-\alpha$ quantile of $F_{p,n-p}$. The dash lines below denote the univariate confidence intervals. The solid circle is the estimated coefficients of age and lbph.



The hypothesis test is given by

$$H_0: \beta_3 = 0, \beta_4 = 0 \qquad H_a: \beta_3 \neq 0 \text{ or } \beta_4 \neq 0.$$

Since the origin lies in the joint 95% confidence interval for β_3 and β_4, the p-value should be larger than 0.05, so the null hypothesis is not rejected.

```r
require(ellipse)
plot(ellipse(model,c(4,5), level=0.95),type="l")
points(coef(model)[4], coef(model)[5], pch=19)
abline(v=confint(model)[4,],lty=2)
abline(h=confint(model)[5,],lty=2)
points(0, 0, pch=1)
```

(c) Suppose a new patient with the following values arrives:

lcavol	lweight	age	Ibph	svi	lcp	gleason	pgg45
1.44692	3.62301	65.00000	0.30010	0.00000	-0.79851	7.00000	15.00000

Predict the lpsa for this patient along with an appropriate 95% CI.

The predicted lpsa of the new sample is 2.3890528 and the corresponding 95% confidence interval is $(0.9646584, 3.813447)$.

Given new sample point $X_{new} \in \mathbb{R}^{p+1}$ (with additional first element being 1), the predicting unkonw observation is

$$Y_{new} = X_{new}^\top \beta + \epsilon_{new},$$

where $\epsilon_{new} \sim N(0, \sigma^2)$. Since

$$\begin{aligned} Y_{new} &\sim N(X_{new}^\top \beta, \sigma^2) \\ \hat{Y}_{new} &\sim N(X_{new}^\top \beta, \sigma^2 X_{new}^\top (X^\top X)^{-1} X_{new}), \end{aligned}$$

we have

$$\frac{\mathbb{E}Y_{new} - \hat{Y}_{new}}{s} \sim t(n-p)$$

where

$$\begin{aligned} s_{pred} &= \sqrt{\hat{\sigma}^2(1 + X_{new}^\top (X^\top X)^{-1} X_{new})} \\ \hat{\sigma}^2 &= \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - X_i^\top \beta)^2. \end{aligned}$$

The $1-\alpha$ confidence interval of Y_{new} is given by $(\hat{Y}_{new} - s_{pred} \cdot t_{n-2}^{(1-\frac{\alpha}{2})}, \hat{Y}_{new} + s_{pred} \cdot t_{n-2}^{(1-\frac{\alpha}{2})})$.

```r
# predict function
new_X <- data.frame(1.44692, 3.62301, 65.00000, 0.30010, 0.00000, -0.79851, 7.00000, 15.00000)
names(new_X) <- colnames(prostate)[-9]
predict(model, new_X, interval='confidence', level=0.95)
```

```
##        fit       lwr      upr
## 1 2.389053 0.9646584 3.813447
```

```r
# raw calculation
n <- nrow(prostate)
X <- cbind(matrix(1, n, 1), as.matrix(prostate)[,-9])
new_X <- t(cbind(1, as.matrix(new_X)))
s_pred <- sqrt(t(new_X) %*% solve(t(X)%*%X) %*% new_X)
sigma_hat <- summary(model)$sigma
df <- model$df.residual
cat('s_pred = ',s_pred)
```

```
## s_pred =  1.231672
```

```r
cat('The 95% prediction interval is (',
    t(new_X) %*% model$coefficients - qt(0.975, df)*sigma_hat*s_pred, ',',
    t(new_X) %*% model$coefficients + qt(0.975, df)*sigma_hat*s_pred, ')')
```

```
## The 95% confidence interval is ( 0.9646584, 3.813447 )
```

(d) Repeat the last question for a patient with the same values except that he or she is age 20. Explain why the CI is wider.

The predicted lpsa of this new sample is 3.2727257 and the corresponding 95% confidence interval is $(2.2604439, 4.2850074)$.

Given new sample point $X'_{new} \in \mathbb{R}^{p+1}$ (with additional first element being 1), the predicting unkonw observation is

$$Y'_{new} = X'^{\top}_{new}\beta + \epsilon'_{new},$$

where $\epsilon'_{new} \sim N(0, \sigma^2)$. Similarly as in (c), the $1-\alpha$ confidence interval of Y'_{new} is given by $(\hat{Y}'_{new} - s_{pred} \cdot t_{n-2}^{(1-\frac{\alpha}{2})}, \hat{Y}'_{new} + s_{pred} \cdot t_{n-2}^{(1-\frac{\alpha}{2})})$.

Notice that the new sample in (c) is just the median sample point of the training set. In (d), the value of age, which is 20, does not lie in the range of ages in the training samples (41,79). Since other covariates are fixed, when the value of age moves far away the average value, it will cause higher estimated variance s. So this CI is wider.

```
# predict function
new_X <- data.frame(1.44692, 3.62301, 20.00000, 0.30010, 0.00000, -0.79851, 7.00000, 15.00000)
names(new_X) <- colnames(prostate)[-9]
predict(model, new_X, interval='prediction', level=0.95)
```

```
##        fit      lwr      upr
## 1 3.272726 1.538744 5.006707
```

```
# raw calculation
n <- nrow(prostate)
X <- cbind(matrix(1, n, 1), as.matrix(prostate)[,-9])
new_X <- t(cbind(1, as.matrix(new_X)))
s_pred <- sqrt(t(new_X) %*% solve(t(X)%*%X) %*% new_X)
sigma_hat <- summary(model)$sigma
df <- model$df.residual
cat('s_pred = ',s_pred)
```

```
## s_pred =  1.231672
```

```
cat('The 95% prediction interval is (',
    t(new_X) %*% model$coefficients - qt(0.975, df)*sigma_hat*s_pred, ',',
    t(new_X) %*% model$coefficients + qt(0.975, df)*sigma_hat*s_pred, ')')
```

```
## The 95% confidence interval is ( 1.538744 , 5.006707 )
```

(e) In the text, we made a permutation test corresponding to the F-test for the significance of all the predictors. Execute the permutation test corresponding to the t-test for age in this model. (Hint: `summary(g)$coef[4,3]` gets you the t-statistic you need if the model is called g.)

The estimated p-value using the permutation test is 0.0855, which is larger than 0.05. So we cannot reject the null hypothesis that $\beta_3 = 0$.

```r
t <- summary(model)$coef[4,3]
df <- model$df.residual
set.seed(0)
n <- 2000
tstats <- numeric(n)
for (i in 1:n){
    ge <- lm(lpsa~lcavol+lweight+sample(age)+lbph+svi+lcp+gleason+pgg45, prostate)
    tstats[i] <- summary(ge)$coef[4,3]
}

mean(abs(tstats) > abs(t))
```

```
## [1] 0.0855
```

3. Using the teengamb data, fit a model with gamble as the response and the other variables as predictors.

(a) Which variables are statistically significant?

The covariates sex and income are statistically significant at the level of $\alpha = 0.05$.

```r
data(teengamb)
model <- lm(gamble~., teengamb)
summary(model)
```

```
##
## Call:
## lm(formula = gamble ~ ., data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

(b) What interpretation should be given to the coefficient for sex?

For sex, 0=male, 1=female. The coefficients of sex is -22.11833, which means that under this linear model, the expenditure on gambling in pounds per year and sex has negative linear correlation, i.e., males tend to spend more expenditure on gambling than females.

(c) Predict the amount that a male with average (given these data) status, income and verbal score would gamble along with an appropriate 95% CI. Repeat the prediction for a male with maximal values (for this data) of status, income and verbal score. Which CI is wider and why is this result expected?

For the first case, the predicted amount is 28.24252 and the corresponding 95% conidence interval is (18.78277, 37.70227). For the second case, the predicted amount is 71.30794 and the corresponding 95% conidence interval is (42.23237, 100.3835). The second CI is wider. First, the value of each predictors except sex is non-negative, and is much larger than the average value. Holding all other variables constant, for the positive semi-definite quadratic form $x^\top(X^\top X)^{-1}x$ for $x \succeq 0$, if the values of some axis get larger enough, it tends to have larger values and so does the estimated variance.

```
new_X <- data.frame(0, mean(teengamb$status), mean(teengamb$income), mean(teengamb$verbal))
names(new_X) <- colnames(teengamb)[-5]
predict(model, new_X, interval='prediction', level=0.95)
```

```
##        fit       lwr      upr
## 1 28.24252 -18.51536 75.00039
```

```
new_X <- data.frame(0, max(teengamb$status), max(teengamb$income), max(teengamb$verbal))
names(new_X) <- colnames(teengamb)[-5]
predict(model, new_X, interval='prediction', level=0.95)
```

```
##        fit      lwr    upr
## 1 71.30794 17.06588 125.55
```

(d) Fit a model with just income as a predictor and use an F-test to compare it to the full model.

$$H_0 : \beta_1 = \beta_2 = \beta_4 = 0, \qquad H_a : \beta_1, \beta_2, \text{ or } \beta_4 \neq 0.$$

Since the p-value of $0.01177 < 0.05$ is so small, this null hypothesis is rejected.

```
reduced_model <- lm(gamble~income, teengamb)
# Anova function
anova(reduced_model, model, test="F")
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ income
## Model 2: gamble ~ sex + status + income + verbal
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     45 28009
## 2     42 21624  3    6384.8 4.1338 0.01177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# F-testing formula
rss0 <- deviance(reduced_model)
rss <- deviance(model)
df0 <- df.residual(reduced_model)
df <- df.residual(model)
fstat <- ((rss0-rss)/(df0-df))/(rss/df)
1-pf(fstat, df0-df, df)
```

```
## [1] 0.01177211
```

4. Suppose that we have a data set following the multiple linear regression model with normal noise,

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i,$$

where $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$. (For simplicity, there's no additional intercept term—as in class, if an intercept is needed then it can be one of the p covariates.) Let $\hat{\beta}$ and $\hat{\sigma}^2$ be the usual estimates of β and σ^2 computed via least squares.
Now let $x^{(0)} \in \mathbb{R}^p$ and $x^{(1)} \in \mathbb{R}^p$ be two new covariate vectors, i.e. you have two new points in your data set, with covariate values $x^{(0)}_1, \ldots, x^{(0)}_p$ for the first new data point and similarly $x^{(1)}_1, \ldots, x^{(1)}_p$ for the second. Let $y^{(0)}$ and $y^{(1)}$ denote the response values for these two data points, which follow the same model, but are unobserved.

(a) What is your estimate for the difference in response values, i.e. for $y^{(0)} - y^{(1)}$?
Since $\hat{y}^{(0)} = x^{(0)\top}\hat{\beta}$, $\hat{y}^{(1)} = x^{(1)\top}\hat{\beta}$, the estimate for the difference in response values is

$$\hat{y}^{(0)} - \hat{y}^{(1)} = (x^{(0)} - x^{(1)})^\top \hat{\beta}.$$

(b) Construct a confidence interval around this estimate with coverage level $1 - \alpha$ (e.g. $\alpha = 0.05$ for 95% confidence).
Since $\hat{\beta} \sim N(\beta, \sigma^2(X^\top X)^{-1})$, we have

$$\hat{y}^{(0)} - \hat{y}^{(1)} = (x^{(0)} - x^{(0)})^\top \hat{\beta} \sim N((x^{(0)} - x^{(1)})^\top \beta, \sigma^2 (x^{(0)} - x^{(1)})^\top (X^\top X)^{-1} (x^{(0)} - x^{(1)})).$$

Also, $y^{(0)} \sim N(x^{(0)\top}\beta, \sigma^2)$, $y^{(1)} \sim N(x^{(1)\top}\beta, \sigma^2)$, so $\mathbb{E}(y^{(0)} - y^{(1)}) = (x^{(0)} - x^{(0)})^\top \beta$. Then

$$\mathbb{E}(y^{(0)} - y^{(1)}) - (\hat{y^{(0)}} - \hat{y^{(1)}}) \sim N(0, \sigma^2 (x^{(0)} - x^{(1)})^\top (X^\top X)^{-1} (x^{(0)} - x^{(1)})),$$

so

$$\frac{\mathbb{E}(y^{(0)} - y^{(1)}) - (\hat{y^{(0)}} - \hat{y^{(1)}})}{s} \sim t_{n-p}$$

where

$$s = \sqrt{\hat{\sigma}^2 (x^{(0)} - x^{(1)})^\top (X^\top X)^{-1} (x^{(0)} - x^{(1)})}.$$

Therefore, the $100 \times (1 - \alpha)\%$ confidence interval is given by $(\hat{y^{(0)}} - \hat{y^{(1)}} - t_{n-p}^{(1-\frac{\alpha}{2})} s, \hat{y^{(0)}} - \hat{y^{(1)}} + t_{n-p}^{(1-\frac{\alpha}{2})} s)$.

(c) Construct a prediction interval for the actual difference $y^{(0)} - y^{(1)}$ with coverage level $1 - \alpha$.
Since $y^{(0)} = x^{(0)\top}\beta + \epsilon^{(0)}$, $y^{(1)} = x^{(1)\top}\beta + \epsilon^{(1)}$, where $\epsilon^{(0)}$ and $\epsilon^{(1)}$ are independent, we have

$$y^{(0)} - y^{(1)} = \epsilon^{(0)} - \epsilon^{(1)} \sim N(0, 2\sigma^2).$$

Also, $\epsilon^{(0)}$ and $\epsilon^{(1)}$ are independent to $\epsilon_1, \ldots, \epsilon_n$, which implies $y^{(0)} - y^{(1)}$ and $\hat{y^{(0)}} - \hat{y^{(1)}}$ are independent. Then

$$(y^{(0)} - y^{(1)}) - (\hat{y^{(0)}} - \hat{y^{(1)}}) \sim N(0, \sigma^2[2 + (x^{(0)} - x^{(1)})^\top (X^\top X)^{-1} (x^{(0)} - x^{(1)})]),$$

so

$$\frac{(y^{(0)} - y^{(1)}) - (\hat{y^{(0)}} - \hat{y^{(1)}})}{s_{pred}} \sim t_{n-p},$$

where

$$s_{pred} = \sqrt{\hat{\sigma}^2 [2 + (x^{(0)} - x^{(1)})^\top (X^\top X)^{-1} (x^{(0)} - x^{(1)})]}.$$

Therefore, the $100 \times (1 - \alpha)\%$ confidence interval is given by $(\hat{y^{(0)}} - \hat{y^{(1)}} - t_{n-p}^{(1-\frac{\alpha}{2})} s_{pred}, \hat{y^{(0)}} - \hat{y^{(1)}} + t_{n-p}^{(1-\frac{\alpha}{2})} s_{pred})$.