STAT 504 Assignment 2

Anuradha Ramachandran

2023-01-29

```r
library('haven')
```

Problem 1

1.1 CEF is the best predictor of Y using X

By decomposition of MSE, for any X=x, $E[(Y-f(x))^2|X=x] = var(y|X=x) + [E[(Y|X=x)] - f(x)]^2$
Differentiating the above equation with f(x) and equating to 0
$\implies E[Y|X=x] = f(x)$, for all $x \in X$
Hence $argmin_f E[(Y-f(x))^2] = E[Y|X]$ will hold for all $x \in X$

1.2 Linear Regression is the best linear predictor of Y using X

Let $f(x) = \alpha + \beta X$, MSE becomes, $E[(Y-(\alpha+\beta X))^2]$
Differentiating MSE with respect to α and equating it to 0
$-2E[Y-\alpha-\beta X] = 0, \implies \alpha = EY - \beta EX$
Differentiating MSE with respect to β and equating it to 0
$-2E[(Y-\alpha-\beta X)X] = 0, \implies \beta = \frac{cov(X,Y)}{var(X)}$
This is same as the coefficients of the linear regression. Hence Linear regression is the best linear predictor of Y using X.

1.3 Linear regression is the best linear approximation of E[Y|X]

Let $Y = E[Y|X], f(x) = \alpha + \beta X$, MSE is given by $E[(E[Y|X] - \alpha - \beta X)^2]$
Differentiating with respect to to α and equating it to 0
$-2E[E[Y|X] - \alpha - \beta X] = 0, \implies \alpha = EY - \beta EX$
Differentiating MSE with respect to β and equating it to 0
$-2E[(E[Y|X] - \alpha - \beta X)X] = 0, \implies \beta = \frac{cov(X,Y)}{var(X)}$
This is same as the coefficients of the linear regression. Hence Linear regression is the best linear approximation of E[Y|X].

We can regress E[Y|X] on X weighted by the probability of X, instead of regressing Y on X.

1.4 Linear CEF

Let CEF, $E[Y|X] = \alpha + \beta X$
For any random variable Y, Y can be expressed as $Y = E[Y|X] + \epsilon = \alpha + \beta X + \epsilon$
$E[Y] = E[\alpha + \beta X + \epsilon] = \alpha + \beta EX + E[\epsilon]$
$E[\epsilon] = 0$ by prop of ϵ
$\alpha = E[Y] - \beta EX$
$cov(Y,X) = cov(\alpha + \beta X + \epsilon, X) = \beta cov(X,X) + cov(X,\epsilon)$
$\beta = \frac{cov(Y,X)}{var(X)}$
Since α and β are similar to the coefficients of BLP, hence if CEF is linear CEF=BLP.

Problem 2

2.1 Curved roof distribution

(a) CEF (E[Y|X]) for $(0 \leq x \leq 2)$

$E[Y|X] = \int_{y=0}^{1} yf(y|x)dy = \int_{y=0}^{1} y\frac{f(x,y)}{f(x)}dy$

$f(x) = \int_{y=0}^{1} f(x,y)dy = 3\frac{(2x^2+1)}{22}$

$E[Y|X] = \int_{y=0}^{1} 2y\frac{x^2+y}{2x^2+1}dy = \frac{1}{3}[\frac{3x^2+2}{2x^2+1}]$ for $(0 \leq x \leq 2)$

(b) BLP of Y given X

$E[XY] = \int_{x=0}^{2}\int_{y=0}^{1} xyf(x,y)dydx = \frac{3}{11}\int_{x=0}^{2}\int_{y=0}^{1} xy(x^2+y)dydx = \frac{8}{11}$

$E[X] = \int_{x=0}^{2} xf(x)dx = \frac{3}{22}\int_{x=0}^{2} x(2x^2+1)dx = 15/11$

$f(y) = \int_{x=0}^{2} f(x,y)dx = \frac{(8+6y)}{11}$

$E[Y] = \int_{y=0}^{1} yf(y)dy = \int_{y=0}^{1} y\frac{(8+6y)}{11}dy = 6/11$

$cov(X,Y) = E[XY] - EXEY = -2/121$

$E[X^2] = \int_{x=0}^{2} x^2 f(x)dx = \frac{3}{22}\int_{x=0}^{2} x^2(2x^2+1)dx = 116/55$

$var(X) = E[X^2] - E[X]^2 = 151/605$

$\beta = \frac{cov(X,Y)}{var(X)} = -10/151$

$\alpha = EY - \beta EX = 96/151$

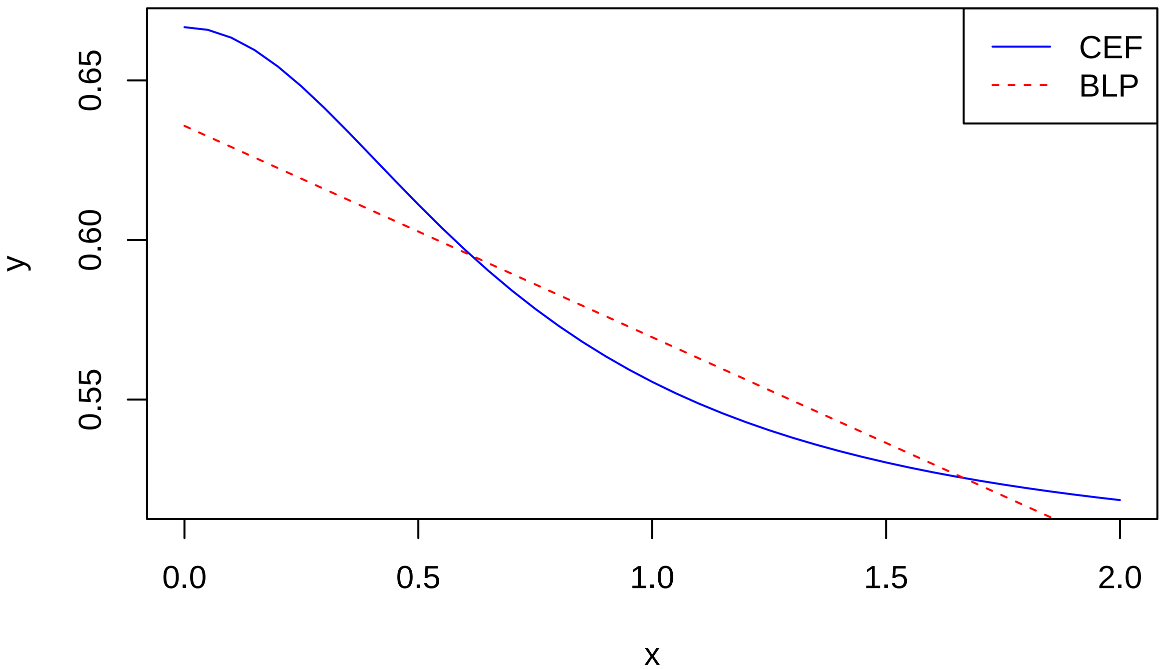
$BLP = \frac{96}{151} - \frac{10}{151}X$

(C) CEF and BLP curve

CEF and the BLP are not the same.

From the plot below, we can see that BLP approximates CEF well in the range of X $(0 \leq X \leq 2)$.

```
x = seq(0,2, by =0.05)
cef = 1/3 *((3*x^2+2)/(2*x^2+1))
blp = (96/151) - (10/151)*x
plot(x,cef, type = 'l', col = "blue", ylab ='y', xlab ='x', lty=1)
points(x,blp, type = 'l', col = "red", lty = 2)
legend(x="topright",legend=c("CEF","BLP"), col=c("blue","red"), lty = c(1,2),ncol=1)
```



2.2 Binary random variable

(a) X and Y are binary random variable. $E[Y|X=1] = \sum_y yP(Y=y|X=1) = P(Y=1|X=1)$
$E[Y|X=0] = P(Y=1|X=0)$
$E[Y|X=x] = E[Y|X=0] + (E[Y|X=1] - E[Y|X=0])X$
$E[Y|X=x] = \alpha + \beta X$ where $\alpha = E[Y|X=0], \beta = E[Y|X=1] - E[Y|X=0]$
Hence E[Y|X] is linear on X.

(b) $var(Y|X) = E[Y^2|X] - E[Y|X]^2$
$E[Y^2|X] = \sum_y y^2 f(y|x)$, is a function of X.
$E[Y|X] = \sum_y yf(y|x)$, is a function of X.
Hence var(Y|X) is a function of X and in general not constant.

(c) As seen in 1.4 if CEF is linear then CEF = BLP. Since CEF is linear on X, linear regression is the best predictor. The regression coefficients meaning:
$\alpha = E[Y|X=0]$ is the expected value of y given x=0
$\beta = E[Y|X=1] - E[Y|X=0]$ is the difference between the expected value of Y given X=1 and X=0.

2.3 Bivariate normal

(a) $E[Y|X] = \mu_y + \sigma_{yx}\sigma_x^{-2}(X - \mu_x)$
$E[X|Y] = \mu_x + \sigma_{xy}\sigma_y^{-2}(Y - \mu_y)$

(b) Let $Y = \alpha_1 + \beta_1 X$
$\beta_1 = \frac{cov(x,y)}{var(x)} = \frac{\sigma_{xy}}{\sigma_x^2}$
$\alpha_1 = \mu_y - \beta_1\mu_x$
$Y = \mu_y - \beta_1\mu_x + \beta_1 X$
$Y = \mu_y + \frac{\sigma_{xy}}{\sigma_x^2}[X - \mu_x] = E[Y|X]$
Hence CEF and linear regressions are the same.

Let $X = \alpha_2 + \beta_2 Y$
$\beta_2 = \frac{cov(x,y)}{var(y)} = \frac{\sigma_{xy}}{\sigma_y^2}$
$\alpha_2 = \mu_x - \beta_2 \mu_y$
$X = \mu_x - \beta_2 \mu_y + \beta_2 Y$
$X = \mu_x + \frac{\sigma_{xy}}{\sigma_y^2}[Y - \mu_y] = E[X|Y]$
Hence CEF and linear regressions are the same.

(c) $var(Y|X) = \sigma_y^2 - \sigma_{xy}\sigma_x^{-2}\sigma_{xy}$
var(Y|X) does not vary as a function of X.

(d) $\beta = E[Y|X = x_0 + 1] - E[Y|X = x_0] = \frac{\partial E[Y|X=x]}{\partial x}|_{X=x}$ is the change in one unit of X that results in change in conditional expectation of Y.
$\alpha = E[Y|X = 0]$ is the conditional expectation of Y given when X=0.

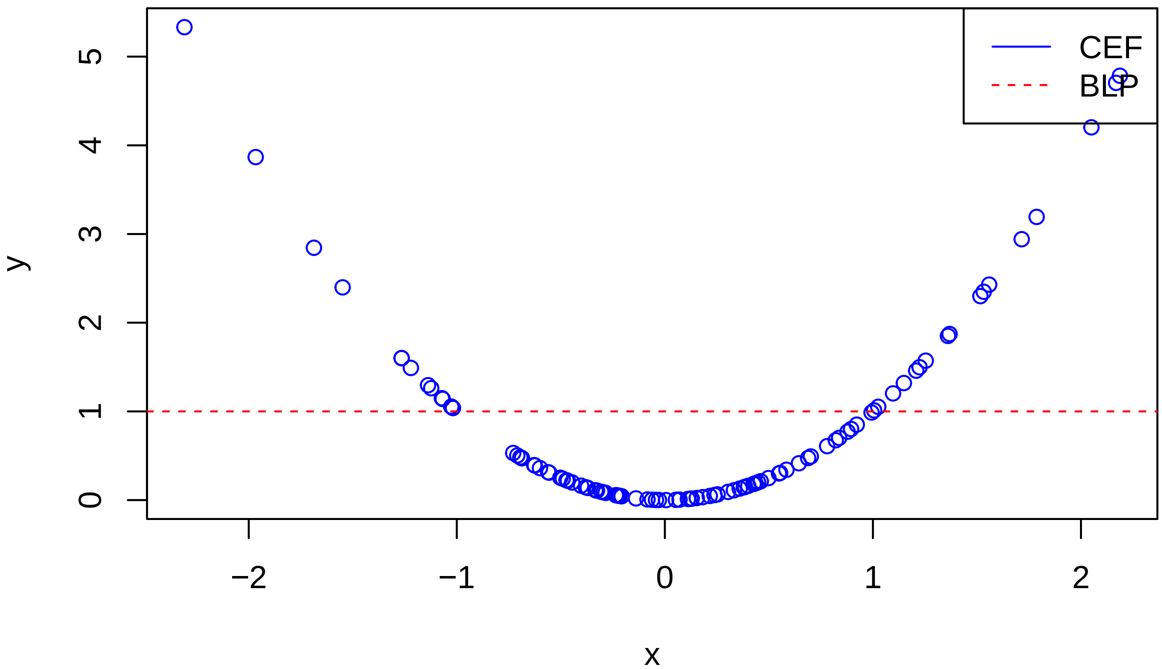
2.4 Quadratic CEF

(a) $E[Y|X] = E[(X^2 + \epsilon)|X] = X^2 + E[\epsilon|X]$
But $E[\epsilon|X] = 0$. Hence $E[Y|X] = X^2$

(b) $\beta = cov(X,Y)/var(X) = \frac{cov(X,X^2+\epsilon)}{var(X)} = \frac{cov(X,X^2)+cov(X,\epsilon)}{var(X)}$
By property of $\epsilon, cov(X,\epsilon) = 0$ and $cov(X,X^2) = 0$
$beta = 0$
$\alpha = EY - \beta EX = EY = E[X^2] + E[\epsilon]$
By property of $E[\epsilon] = 0$
$\alpha = E[X^2] = var(X) + E[X]^2 = 1$
$BLP = \alpha + \beta X = 1$

(c) The CEF and the BLP curves are very different. The BLP of Y on X does not provide a good approximation to the quadratic CEF.

```
set.seed(123)
x = rnorm(100, 0, 1)
epsi = rnorm(100,0,1)
y= x*x + epsi
blp = 1
plot(x*x~x, type ='p', col = "blue", ylab ='y', xlab ='x', lty=1)
abline(blp,0, col = "red", lty = 2)
legend(x="topright",legend=c("CEF","BLP"), col=c("blue","red"), lty = c(1,2),ncol=1)
```



(d)
In this case,
$Q = E[Y|X = a] - E[Y|X = -a] = a^2 - a^2 = 0$
$Q' = BLP[Y|X = a] - BLP[Y|X = -a] = 1 - 1 = 0$
Hence $Q = Q'$

(e) If distribution of X is changed, the CEF will still be equal to X^2.
If distribution of X is changed, the BLP will change since $\alpha and \beta$ depends on the distribution of X.
$\alpha = E[X^2] = var(X) + E[X]^2$ and
$\beta = \frac{cov(X,X^2))}{var(X)}$

(f)
Let $BLP[Y|X^2] = Y = \alpha + \beta X^2 + e$
$E[Y] = \alpha + \beta E[X^2] + E[e]$
Y is defined as $X^2 + \epsilon$
$E[e] = E[\epsilon] = 0$
$\beta = \frac{cov(Y,X^2)}{var(X^2)}$
$\implies cov(\epsilon, X^2) = 0$
$\beta = \frac{cov(X^2,X^2)}{var(X^2)} = 1$
$\alpha = EY - \beta E[X^2] = E[X^2] - E[X^2] = 0$
$BLP[Y|X^2] = X^2$
Hence $BLP[Y|X^2] = CEF$

Problem 3

3.1 CEF decomposition

$Y = f(X) + \epsilon$

Let $f(X) = E[Y|X]$

(a)
$E[\epsilon|X] = E[(Y - E[Y|X])|X] = E[Y|X] - E[Y|X] = 0$
$E[\epsilon] = E[Y - E[Y|X]] = EY - EY = 0$

(b)
$var(\epsilon|X) = E[(\epsilon - E[\epsilon|X])^2|X] = E[\epsilon^2|X] = E[(Y - E[Y|X])^2|X] = var(Y|X)$
$var(\epsilon) = E[var(\epsilon|X)] + var(E[\epsilon|X]) = E[var(\epsilon|X)] = E[var(Y|X)]$

(c)
$E[h(X)\epsilon] = E[E[h(X)\epsilon|X]] = E[h(X)E[\epsilon|X]] = 0$

$f(x) = E[Y|X]$ is the conditional distribution of Y given X
ϵ is the unpredictable error in predicting Y from conditional expectation function.

As proven above these are the properties of the error term ϵ.

3.2 OLS decomposition

(a)
$E[e] = E[Y - \alpha - \beta X] = E[Y] - E[Y] + \beta E[X] - \beta E[X] = 0$

(b)
$E[Xe] = E[X(Y - \alpha - \beta X)] = E[XY - \alpha X - \beta X^2] = E[XY - X(\alpha + \beta X)] = E[XY - XY] = 0$

(c)
$var(Y) = var(\alpha + \beta X + e) = \beta^2 var(X) + var(e)$
$var(e) = var(Y) - \beta^2 var(X)$

$\alpha + \beta X$ is the linear regression function of Y.
e is the error due to non linearity. As proven above these are the properties of the error term e.
β is the coefficient of X in the linear regression model and describes the change in y due to change in x.

3.3

$f(X) \neq \alpha + \beta X$, since the properties of the error terms are different. ϵ measures the irreducible error in predicting Y using the CEF. e measures the non-linearity in the predicting Y using BLP.

3.4

(a) Linear predictor of Y
(b) Difference between the CEF and BLP predictor.
(c) ϵ is the irreducible error (noise) in predicting Y using the CEF

3.5

The book says that the $E[\epsilon|X] = 0$ is an assumption of the error term. However as we have shown above that $E[\epsilon|X] = 0$ is the property of the error terms and not an assumption.

Extra credit

The following passage was obtained from Statistics for machine learning by Pratap Dangeti [Pg#58]. Here, error terms should have zero mean is provided as an assumption, however it is a property of the linear model.

```r
knitr::include_graphics("/Users/anuram/Downloads/IMG_4143.png")
```

Assumptions of linear regression

Linear regression has the following assumptions, failing which the linear regression model does not hold true:

- The dependent variable should be a linear combination of independent variables
- No autocorrelation in error terms
- Errors should have zero mean and be normally distributed
- No or little multi-collinearity
- Error terms should be homoscedastic

Problem 4

4.1

Let $Y = E[Y|X] + \epsilon$
$var(Y) = var(E[Y|X] + \epsilon) = var(E[Y|X]) + var(\epsilon) + 2cov(E[Y|X], \epsilon)$
$var(\epsilon) = E[var(Y|X)]$
$var(Y) = var(E[Y|X]) + E[var(Y|X)]$

4.2

(a) $\eta^2_{y\sim x} = \frac{var(Y) - E[var(Y|X)]}{var(Y)} = 1 - \frac{E[var(Y|X)]}{var(Y)}$
$cor^2(Y, E[Y|X]) = \frac{cov(Y, E[Y|X])^2}{var(Y)var(E[Y|X])} = \frac{[cov(E[Y|X], E[Y|X]) + cov(\epsilon, Z)]^2}{var(Y)var(E[Y|X])} = \frac{var(E[Y|X])}{var(Y)} = \eta^2_{y\sim x}$

(b) In general, $\eta^2_{y\sim x} \neq \eta^2_{x\sim y}$, It is symmetric when $Y = E[Y|X] + \epsilon$ is invertible.
Let $Y = E[Y|X] + \epsilon$, where $E[Y|X] = X^2$
$\eta^2_{y\sim x} = cor^2(Y, X^2) = \frac{cov(Y, X^2)^2}{var(Y)var(X^2)}$
$\eta^2_{x\sim y} = cor^2(X, \pm\sqrt{Y - \epsilon}) = \frac{cov(X, \pm\sqrt{Y-\epsilon})^2}{var(X)var(\pm\sqrt{Y-\epsilon})}$
Hence $\eta^2_{y\sim x} \neq \eta^2_{x\sim y}$

4.3

$var(Y) = var(\alpha + \beta X + e) = var(\alpha + \beta X) + var(e) + 2cov(\alpha + \beta X, e)$
since $cov(\alpha + \beta X, e) = 0$
$var(Y) = var(\alpha + \beta X) + var(e)$

4.4

(a)
$R^2_{y\sim x} = \frac{var(Y) - var(e)}{var(Y)} = 1 - \frac{var(e)}{var(Y)}$
$cor^2(Y, \alpha + \beta X) = \frac{[cov(Y, \alpha+\beta X)]^2}{var(Y)var(\alpha+\beta X)} = \frac{[cov(\alpha+\beta X, \alpha+\beta X)]^2}{var(Y)var(\alpha+\beta X)} = \frac{var(\alpha+\beta X)}{var(Y)}$
$cor^2(Y, \alpha + \beta X) = \frac{[cov(Y, \alpha+\beta X)]^2}{var(Y)var(\alpha+\beta X)} = \frac{\beta^2[cov(Y,X)]^2}{\beta^2 var(Y)var(X)} = cor^2(Y, X)$

(b)
From (a) $R^2_{y\sim x} = cor^2(Y,X) = \frac{cov(Y,X)^2}{var(Y)var(X)}$
By symmetry property of covariance, $cov(Y,X) - cov(X,Y)$
$\implies, \frac{cov(Y,X)^2}{var(Y)var(X)} = \frac{cov(X,Y)^2}{var(Y)var(X)} = cor^2(X,Y) = R^2_{x\sim y}$

4.5

(a)
From the definition of $R^2_{y\sim x} = 1 - \frac{var(e)}{var(Y)} \in [0,1]$ since $var(e) \geq 0$, $R^2_{y\sim x} \leq 1$.
From the definition of $\eta^2_{y\sim x} = 1 - \frac{E[var(Y|X)]}{var(Y)} \in [0,1]$ since $\frac{E[var(Y|X)]}{var(Y)} \geq 0$, $\eta^2_{y\sim x} \leq 1$

Consider $var(E[Y|X])$. Y can also be written as $Y = \alpha + \beta X + e$
$var(E[Y|X]) = var(E[\alpha + \beta X + e]|X) = var(\alpha + \beta X) + var(E[e|X]) + 2cov(\alpha + \beta X, E[e|X])$
$\implies var(E[Y|X]) \geq var(\alpha + \beta X)$
$\implies \eta^2_{y\sim x} \geq R^2_{y\sim x}$

(b) Consider $\eta^2_{y\sim x} - R^2_{y\sim x} = \frac{var(E[Y|X]) - var(\alpha+\beta X)}{var(Y)} = \frac{var(e)}{var(Y)} = \frac{var(e)(var(u))^2}{var(Y)var(e)var(u)} = \frac{var(e)(cov(u,u))^2}{var(Y)var(e)var(u)}$
$\eta^2_{y\sim x} - R^2_{y\sim x} = \frac{var(e)(cov(u,Y-\alpha-\beta X-\epsilon))^2}{var(Y)var(e)var(u)}$
$\implies \frac{var(e)(cov(u,Y-\alpha-\beta X))^2}{var(Y)var(e)var(u)} = \frac{var(e)(cov(u,e))^2}{var(Y)var(e)var(u)} = \frac{var(e)}{var(Y)}cor^2(u,e) = (1 - R^2_{y\sim x})cor^2(u,e)$
$\eta^2_{y\sim x} = R^2_{y\sim x} + (1 - R^2_{y\sim x})cor^2(u,e)$

Problem 5

5.1 Linear transformation of X

Let $Y = \alpha + \beta X + e$
Let $X' = a + bX$
$Y = \alpha + \beta\frac{(X'-a)}{b} + e = \alpha + \frac{\beta}{b}X' - \frac{\beta a}{b} + e$
$\alpha' = \alpha - \frac{\beta a}{b}$, $\beta' = \frac{\beta}{b}$, $e' = e$
Hence $Y = \alpha' + \beta' X' + e'$
$cor^2(Y,X') = cor^2(Y, a+bX) = \frac{(cov(Y,a+bX))^2}{var(Y)var(a+bX)} = \frac{b^2(cov(Y,X))^2}{b^2 var(Y)var(X)} = cor^2(Y,X) = R^2_{y\sim x}$
Hence R^2 remains unchanged due to this transformation of X.

5.2 Linear transformation of Y

Let $Y = \alpha + \beta X + e$
Let $Y' = a + bY$
$\frac{Y'-a}{b} = \alpha + \beta X + e$
$Y' = b\alpha + b\beta X + eb + a$
$\alpha' = a + b\alpha$, $\beta' = \beta b$, $e' = eb$
$cor^2(Y',X) = cor^2(a+bY, X) = \frac{(cov(a+bY,X))^2}{var(a+bY)var(X)} = \frac{b^2(cov(Y,X))^2}{b^2 var(Y)var(X)} = cor^2(Y,X) = R^2_{y\sim x}$
Hence R^2 remains unchanged due to this transformation of Y.

5.3 Standarization

Let $Y = \alpha + \beta X + e$
Let $X' = \frac{X-EX}{SD(X)}$ and $Y' = \frac{Y-EY}{SD(Y)}$

(a) E[X'] and E[Y']
$E[X'] = E[\frac{X-EX}{SD(X)}] = \frac{1}{SD(X)}(E[X] - E[X]) = 0$
$E[Y'] = E[\frac{Y-EY}{SD(Y)}] = \frac{1}{SD(Y)}(E[Y] - E[Y]) = 0$

(b) var[X'] and var[Y']
$var(X') = E[X'^2] - E[X']^2 = E[X'^2] = \frac{1}{SD(X)^2}(E[(X - E[X])^2]) = \frac{var(X)}{SD(X)^2} = 1$
$var(Y') = E[Y'^2] - E[Y']^2 = E[Y'^2] = \frac{1}{SD(Y)^2}(E[(Y - E[Y])^2]) = \frac{var(Y)}{SD(Y)^2} = 1$

(c) transformation
Since $Y' = \alpha' + \beta' X + e'$
$\alpha' = E[Y'] - \beta' E[X'] = 0$
$\beta' = \frac{cov(X',Y')}{var X'} = \frac{E[X'Y'] - E[X']E[Y']}{var(X')} = E[X'Y']$
$\implies \beta' = E[(\frac{(X-EX)}{SD(X)})(\frac{(Y-EY)}{SD(Y)})] = E[\frac{1}{SD(X)SD(Y)}(X - EX)(Y - EY)] = \frac{cov(X,Y)}{SD(X)SD(Y)} = \beta\frac{SD(X)}{SD(Y)}$
$cor^2(Y', X') = \frac{cov(Y',X')^2}{var(Y')var(X')} = cov(Y', X')^2$
$cor^2(Y', X') = \frac{cov(Y-EY,X-EX)^2}{var(Y)var(X)} = \frac{cov(Y,X)^2}{var(Y)var(X)} = cor^2(Y, X) = R^2_{y\sim x}$
Hence R^2 remains unchanged due to this transformation of Y and X.

Problem 6

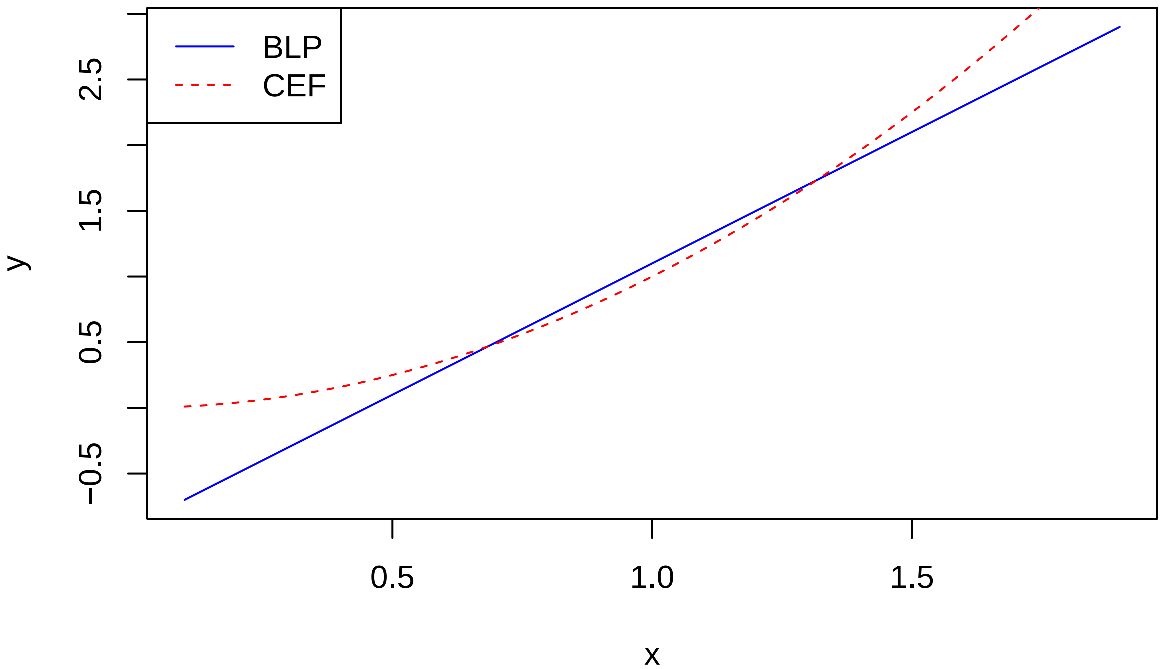
$\epsilon \sim N(0, 1)$
$Y = X^2 + \epsilon$

6(a)

$X \sim N(1, 0.1)$
$CEF = E[Y|X] = E[X^2 + \epsilon|X] = X^2$
BLP coefficients:
$\beta = \frac{cov(X,Y)}{var(X)} = \frac{cov(X,X^2)}{0.1} = \frac{E[X^3]-E[X]E[X^2]}{var(X)}$
$E[X^2] = var(X) + E[X]^2 = 1.1$
$E[X^3] = E[X]^3 + 3E[X]var[X] = 1.3$
$\beta = \frac{1.3-1.1}{0.1} = 2$
$\alpha = EY - \beta EX = E[X^2] - \beta EX = 1.1 - 2 = -0.9$ $BLP = -0.9 + 2X$

From the plot below, the BLP provides a good estimate between approximately when X= 0.5 to 1.5, however CEF and BLP both diverge towards the end of the interval (0.1,1.9)

```
x = seq(from = 0.1, to=1.9,length = 100)
cef = x*x
blp = -0.9 + 2*x
plot(x,blp, type='l', col = "blue", ylab = "y", xlab = "x", lty = 1)
lines(x, cef, col = "red",type='l', lty=2)
legend(x="topleft",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(1,2),ncol=1)
```



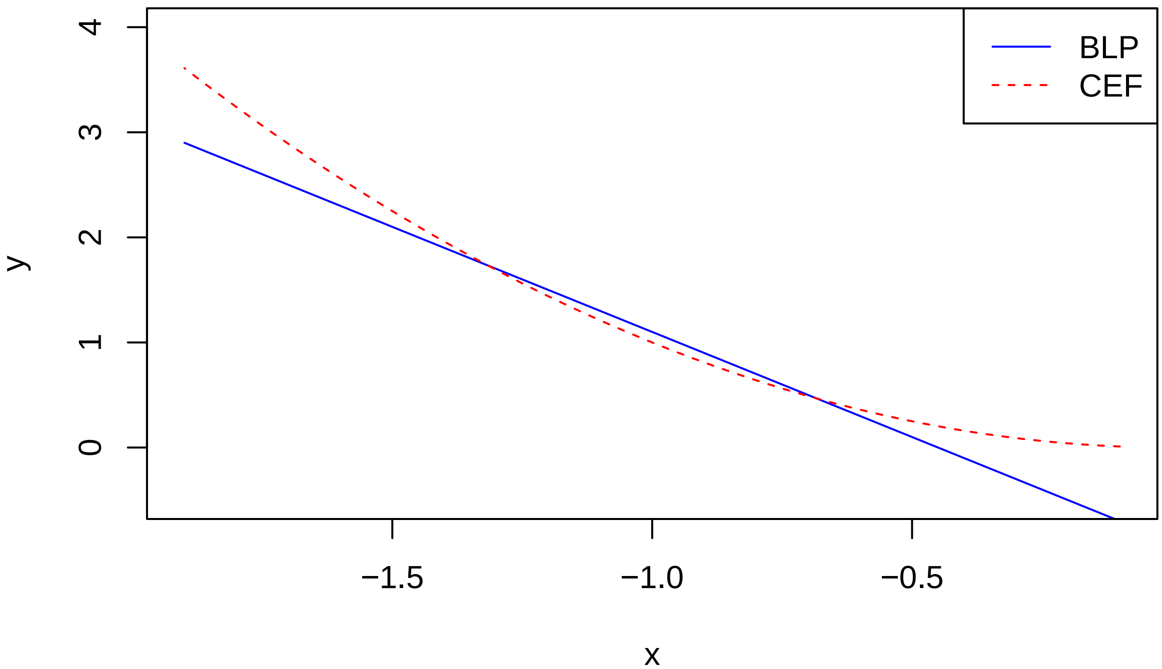
6(b)

$X \sim N(-1, 0.1)$
$CEF = E[Y|X] = E[X^2 + \epsilon|X] = X^2$
BLP coefficients:
$\beta = \frac{cov(X,Y)}{var(X)} = \frac{cov(X,X^2)}{0.1} = \frac{E[X^3]-E[X]E[X^2]}{var(X)}$
$E[X^2] = var(X) + E[X]^2 = 1.1$
$E[X^3] = E[X]^3 + 3E[X]var[X] = -1.3$
$\beta = \frac{-1.3+1.1}{0.1} = -2$
$\alpha = EY - \beta EX = E[X^2] - \beta EX = 1.1 - (-2) * (-1) = -0.9$ $BLP = -0.9 - 2X$

From the plot below, the BLP provides a good estimate between approximately when X= -1.5 to -0.5, however CEF and BLP both diverge towards the end of the interval (-1.9,-0.1)

The BLP in (b) is not same as the BLP in (a).

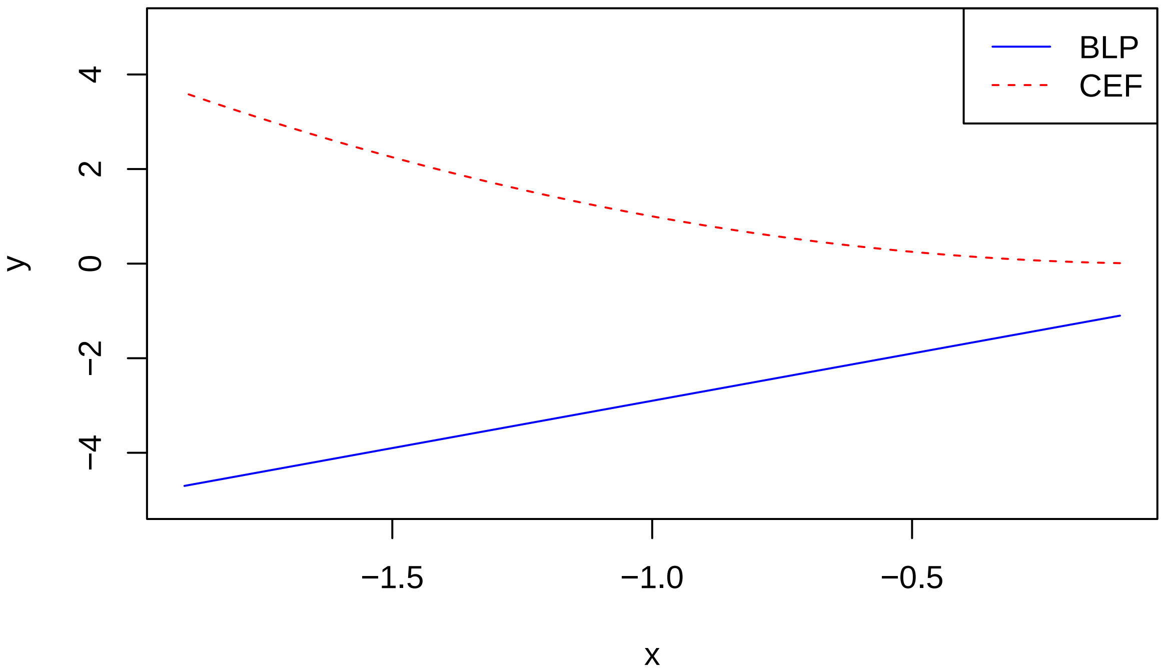
```
x = seq(from = -0.1, to= -1.9,length = 100)
cef = x*x
blp = -0.9 - 2*x
plot(x,blp, type='l', ylim = c(-0.5, 4),col = "blue", ylab = "y", xlab = "x", lty = 1)
lines(x, cef, col = "red",type='l', lty=2)
legend(x="topright",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(1,2),ncol=1)
```



6(c)

The blp from (a) does not provide a good prediction of the CEF in (b). The blp from (a) was trained on a different distribution of $X \sim N(1, 0.1)$ with X ranging from [0.1, 1.9]. Hence it does not do a good job of predicting the CEF from a different distribution of $X \sim N(-1, 0.1)$ in the range of X = (-1.9,-0.1).

```
x = seq(from = -0.1, to= -1.9,length = 100)
cef = x*x
blp = -0.9 + 2*x
plot(x,blp,ylim = c(-5,5), type='l', col = "blue", ylab = "y", xlab = "x", lty = 1)
lines(x, cef, col = "red",type='l', lty=2)
legend(x="topright",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(1,2),ncol=1)
```



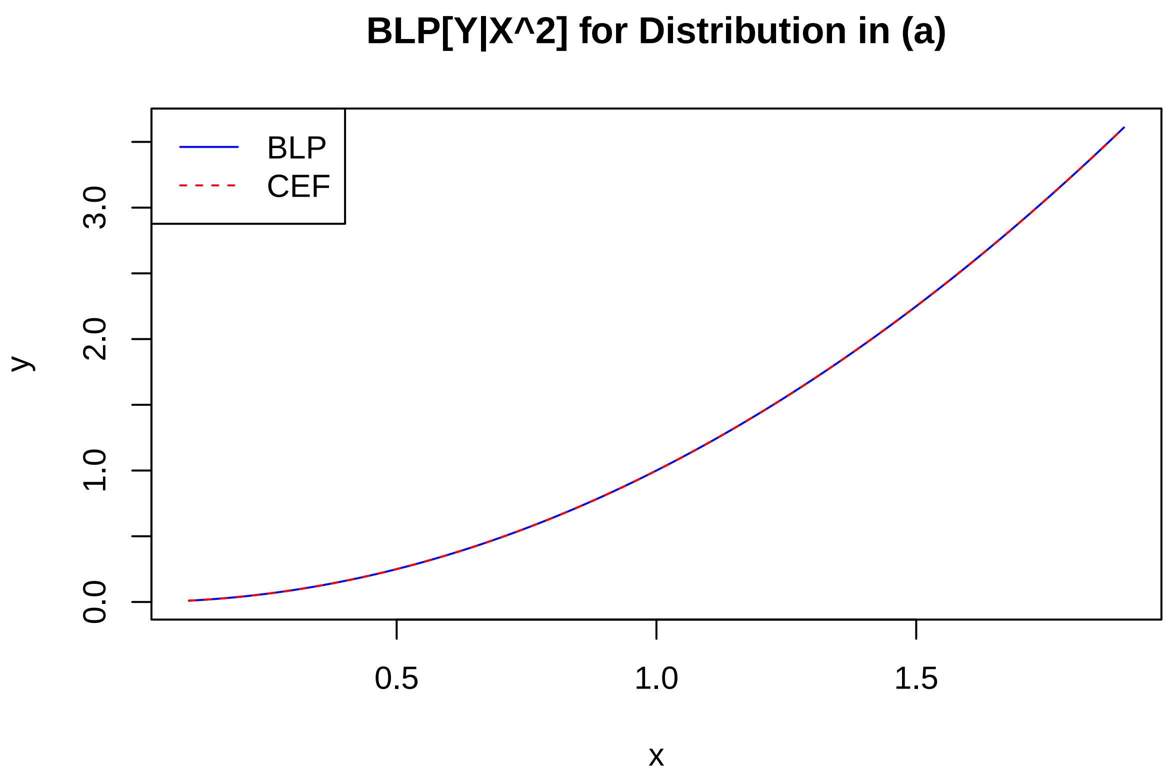
6(d)

$CEF = E[Y|X] = E[X^2 + \epsilon|X] = X^2$
$BLP[Y|X^2]$ BLP coefficients:
$\beta = \frac{cov(X^2,Y)}{var(X)} = \frac{cov(X^2,X^2)}{var(X)} = 1$ for both distributions of X in (a) and (b) since the variance in both cases are the same.

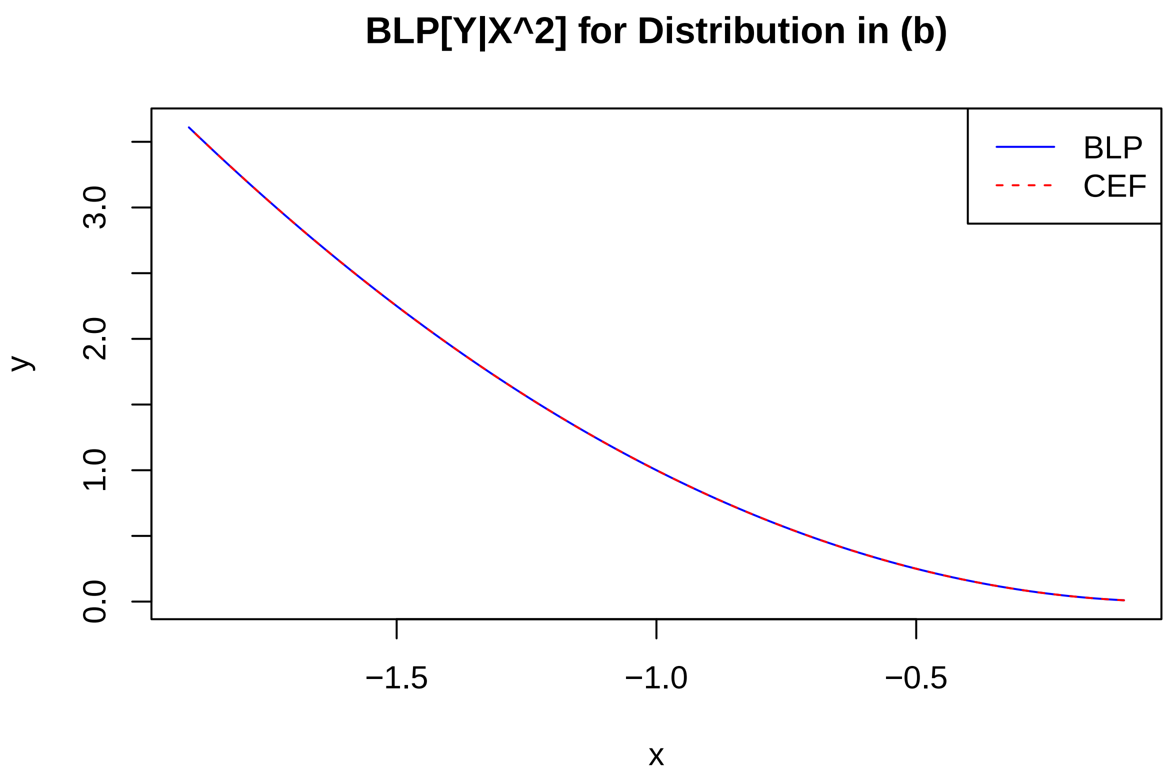
Let the BLP model be $Y = \alpha + \beta X^2 + e$
$E[Y] = \alpha + \beta E[X^2]$
$\implies \alpha = E[Y] - \beta E[X^2] = E[X^2 + \epsilon] - E[X^2] = E[X^2] - E[X^2] = 0$
For both (a) and (b), the BLP is $Y = X^2$

The $BLP[Y|X^2]$ fits the CEF perfectly in all the cases (a), (b) and (c)

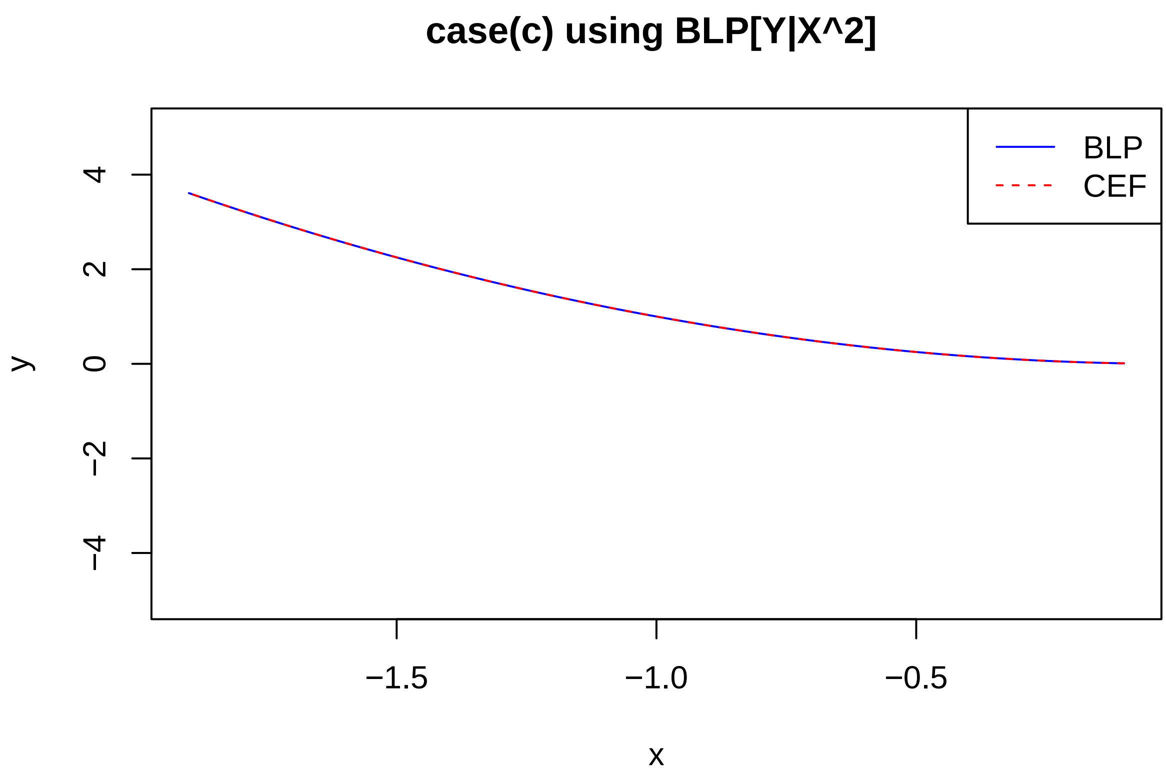
```
x = seq(from = 0.1, to= 1.9,length = 100)
cef = x*x
blp = x*x
plot(x,blp,  type='l', col = "blue", ylab = "y", xlab = "x", main = "BLP[Y|X^2] for Distribution in (a)
lines(x, cef, col = "red",type='l', lty=2)
legend(x="topleft",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(1,2),ncol=1)
```




```r
x = seq(from = -0.1, to= -1.9,length = 100)
cef = x*x
blp = x*x
plot(x,blp,  type='l', col = "blue", ylab = "y", xlab = "x", main = "BLP[Y|X^2] for Distribution in (b)
lines(x, cef, col = "red",type='l', lty=2)
legend(x="topright",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(1,2),ncol=1)
```

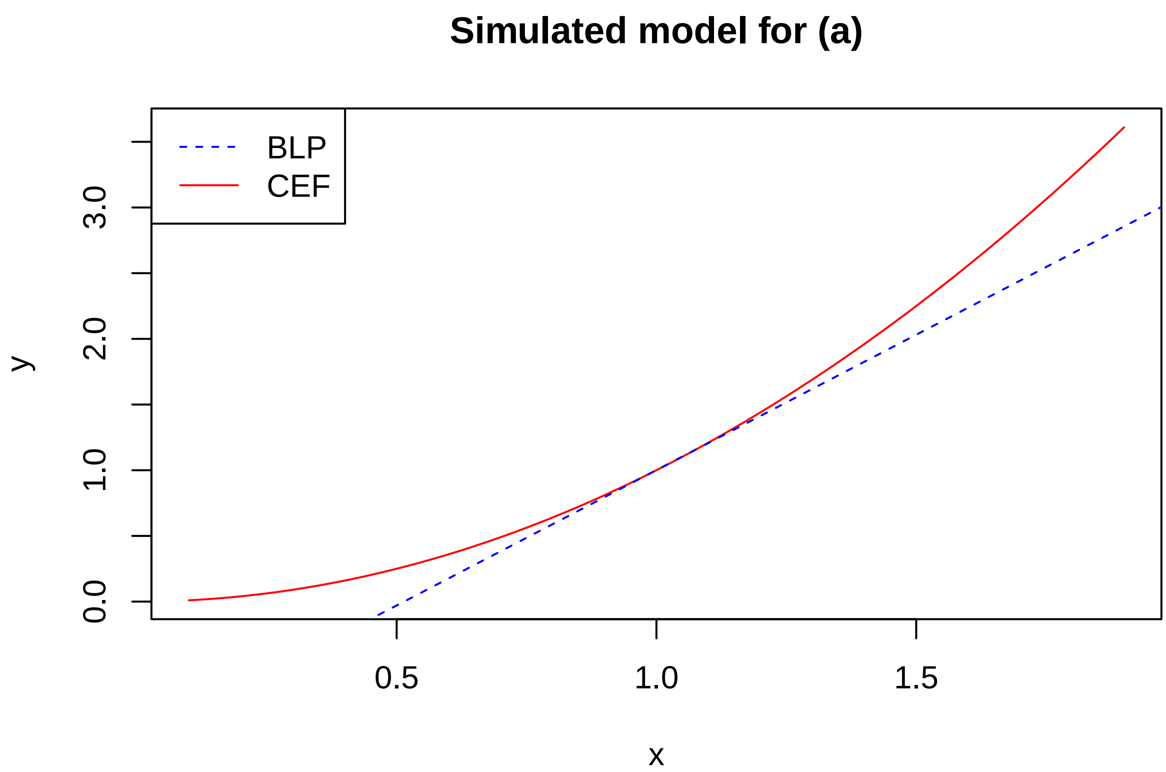



```
x = seq(from = -0.1, to= -1.9,length = 100)
cef = x*x
blp = x*x
plot(x,blp,ylim = c(-5,5), type='l', col = "blue", ylab = "y", xlab = "x",main = "case(c) using BLP[Y|X
lines(x, cef, col = "red",type='l', lty=2)
legend(x="topright",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(1,2),ncol=1)
```

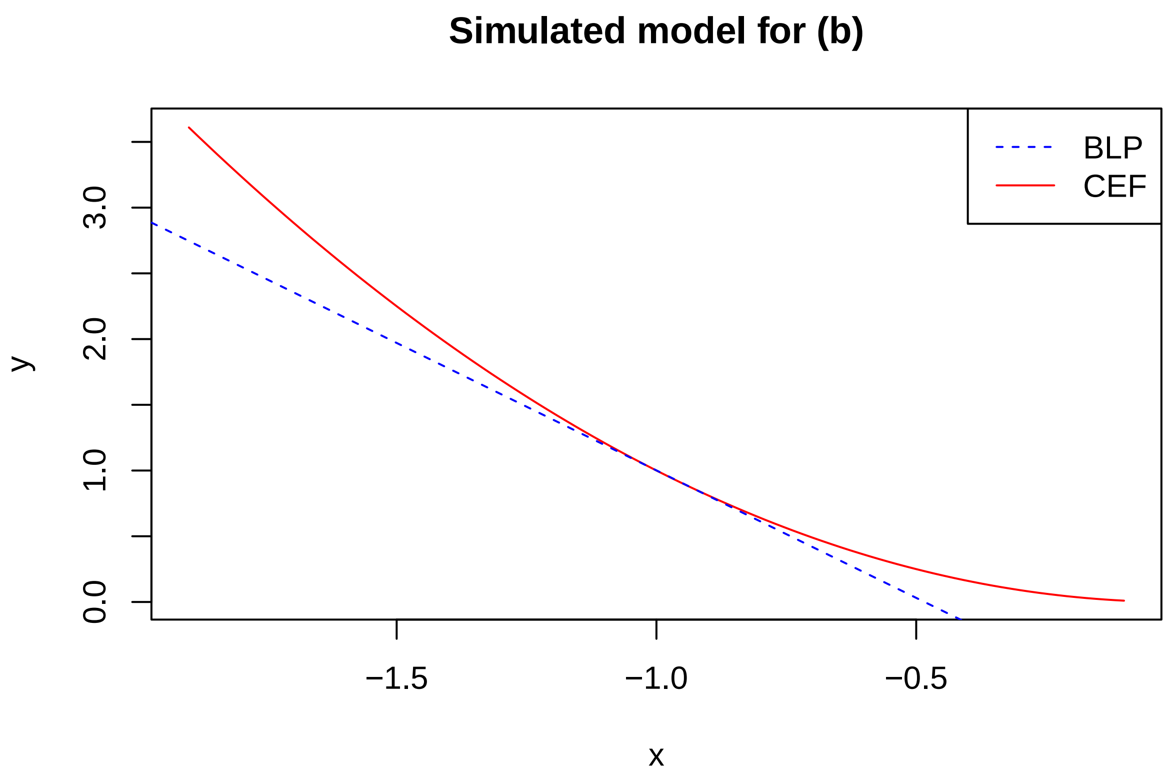


6(e)

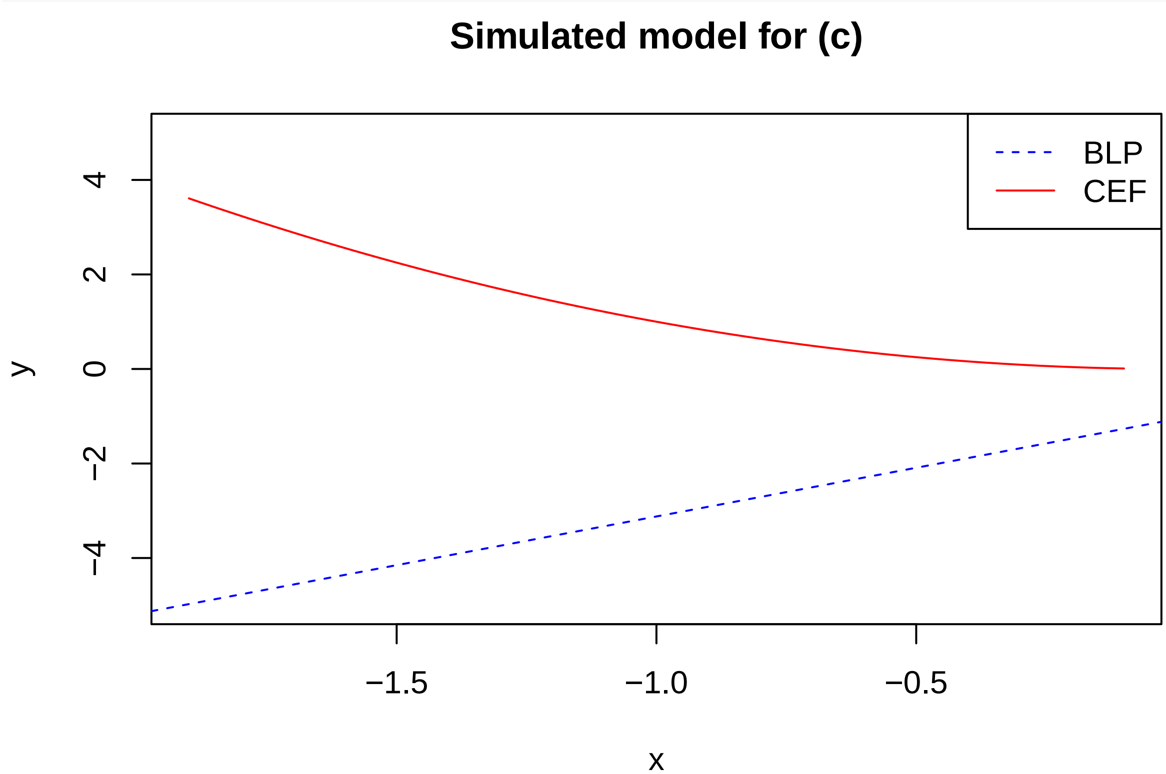
```r
set.seed(123)
x = rnorm(10000, 1, 0.1)
epsi = rnorm(10000, 0,1)
y = x*x + epsi
blp_a = lm(y~x)
xlim = seq(from = 0.1, to= 1.9,length = 100)
cef = xlim*xlim
plot(xlim,cef, type='l', col = "red", ylab = "y", xlab = "x", main = "Simulated model for (a)", lty = 1
abline(blp_a, col = "blue", lty=2)
legend(x="topleft",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(2,1),ncol=1)
```



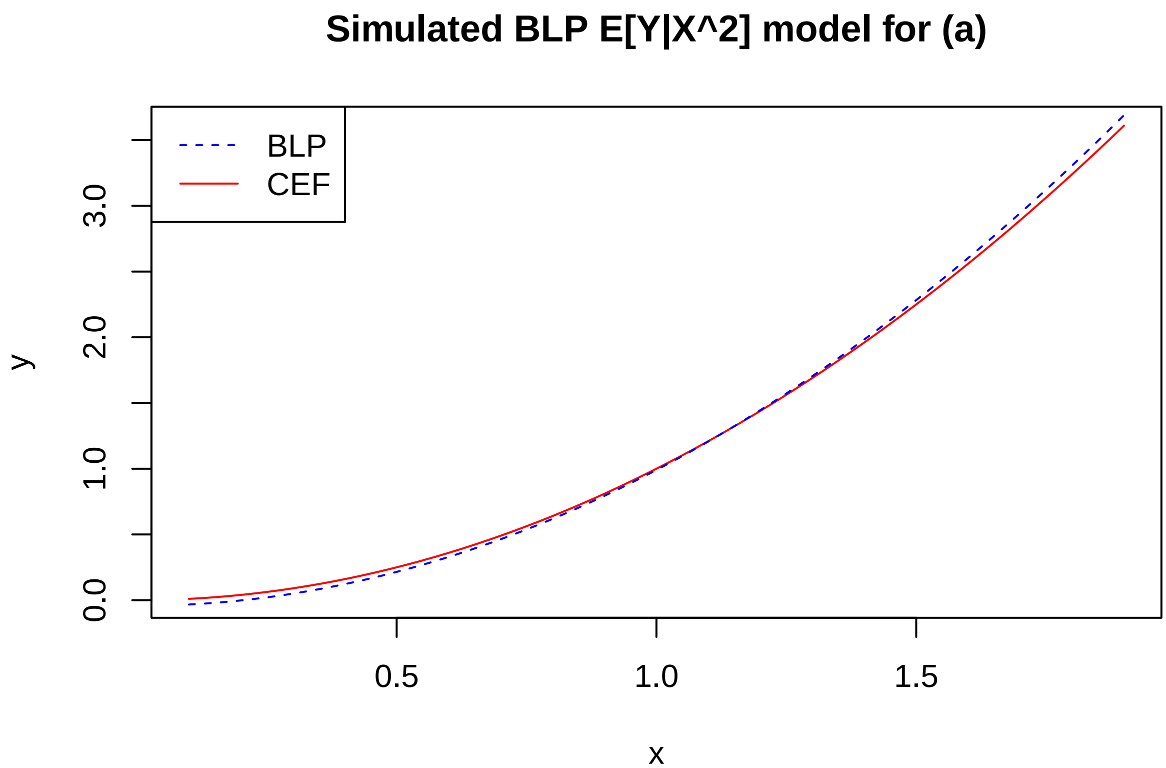

```r
set.seed(123)
x = rnorm(10000, -1, 0.1)
epsi = rnorm(10000, 0,1)
y = x*x + epsi
blp = lm(y~x)
xlim = seq(from = -0.1, to= -1.9,length = 100)
cef = xlim*xlim
plot(xlim,cef, type='l', col = "red", ylab = "y", xlab = "x", main = "Simulated model for (b)", lty = 1
abline(blp, col = "blue", lty=2)
legend(x="topright",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(2,1),ncol=1)
```



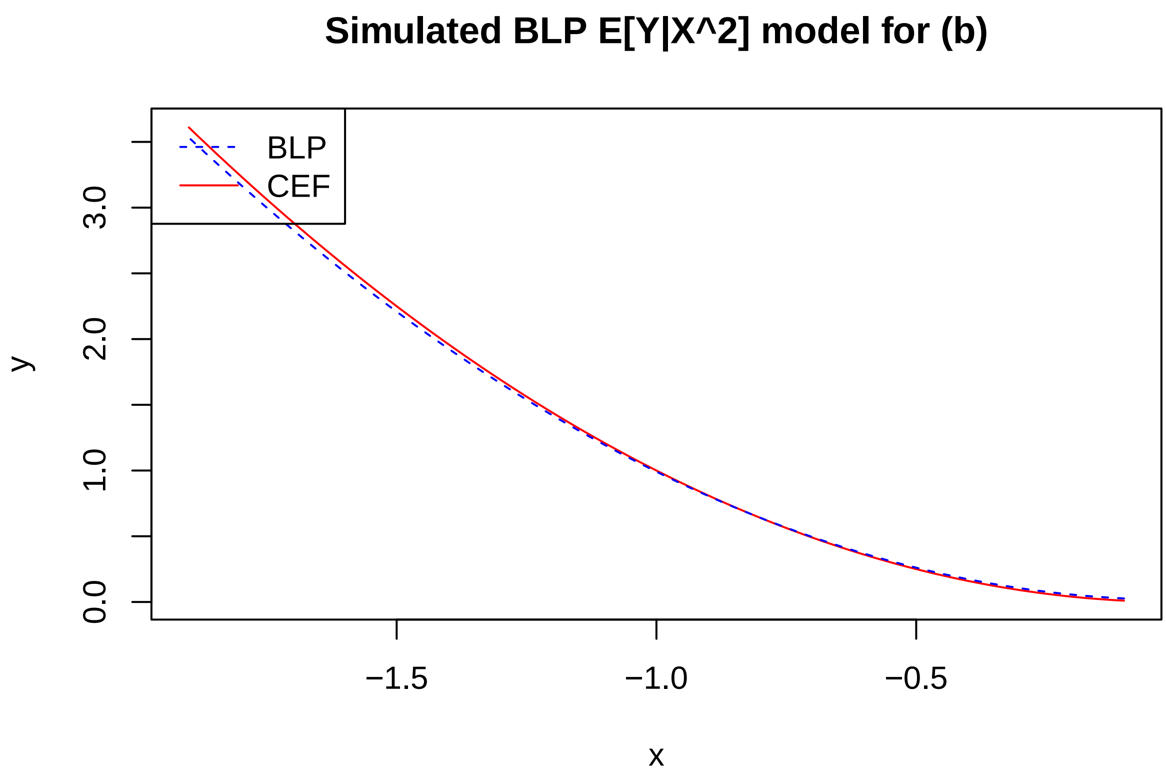

```
x = seq(from = -0.1, to= -1.9,length = 100)
cef = x*x
plot(x,cef, type='l', ylim = c(-5,5), col = "red", ylab = "y", xlab = "x", main = "Simulated model for
abline(blp_a, col = "blue", lty=2)
legend(x="topright",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(2,1),ncol=1)
```



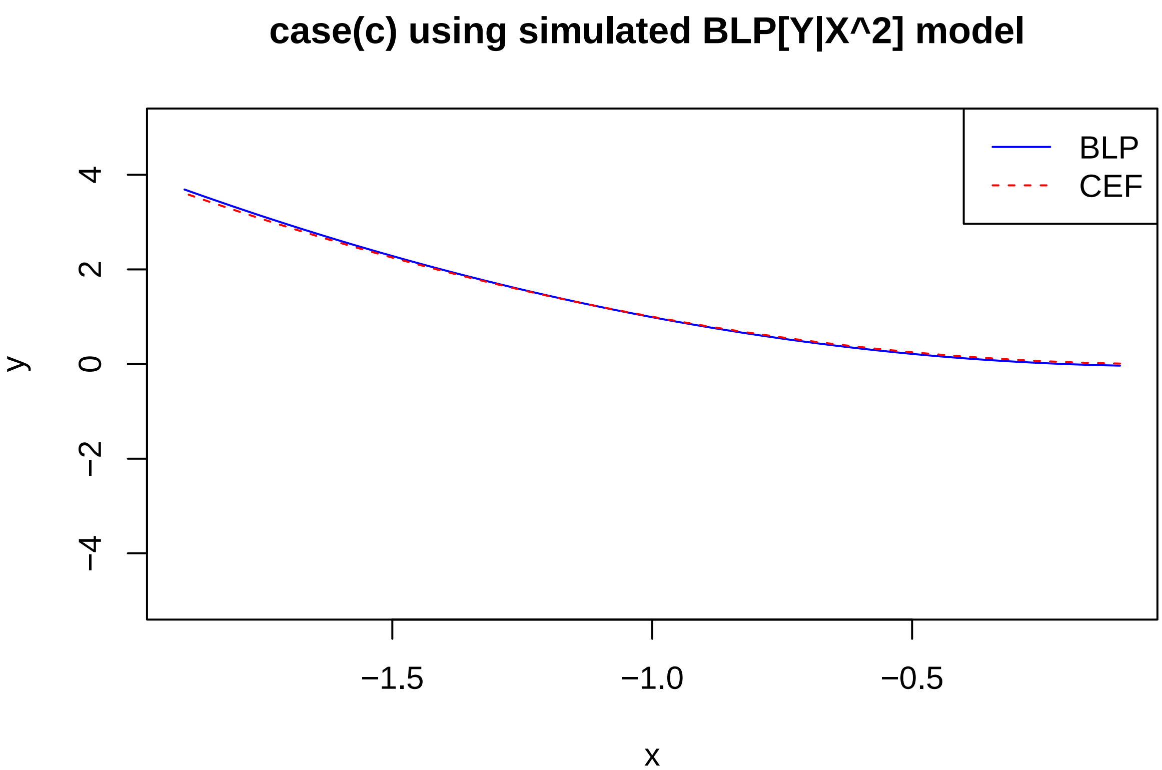
```
set.seed(123)
x = rnorm(10000, 1, 0.1)
epsi = rnorm(10000, 0,1)
y = x*x + epsi
x_2 = x*x
blp_a = lm(y~x_2)
xlim = seq(from = 0.1, to= 1.9,length = 100)
cef = xlim*xlim
blp = predict(blp_a, list(x_2 = xlim^2))
plot(xlim,cef, type='l', col = "red", ylab = "y", xlab = "x", main = "Simulated BLP E[Y|X^2] model for
lines(xlim, blp,type='l',col = "blue", lty=2)
legend(x="topleft",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(2,1),ncol=1)
```




```
set.seed(123)
x = rnorm(10000, -1, 0.1)
epsi = rnorm(10000, 0,1)
y = x*x + epsi
x_2 = x*x
blp_b= lm(y~x_2)
xlim = seq(from = -0.1, to= -1.9,length = 100)
cef = xlim*xlim
blp = predict(blp_b, list(x_2 = xlim^2))
plot(xlim,cef, type='l', col = "red", ylab = "y", xlab = "x", main = "Simulated BLP E[Y|X^2] model for
lines(xlim, blp,type='l',col = "blue", lty=2)
legend(x="topleft",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(2,1),ncol=1)
```




```
x = seq(from = -0.1, to= -1.9,length = 100)
cef = x*x
blp = predict(blp_a, list(x_2 = x^2))
plot(x,blp,ylim = c(-5,5), type='l', col = "blue", ylab = "y", xlab = "x",main = "case(c) using simulat
lines(x, cef, col = "red",type='l', lty=2)
legend(x="topright",legend=c("BLP","CEF"), col=c("blue","red"),lty=c(1,2),ncol=1)
```



Problem 7

7(a)

The linear regression of call on black model coefficients are:
Intercept: 0.096509
Slope: -0.032033

Intercept: If the name of the candidate is not "black sounding" ie: when the variable "black" is 0, then on average the callback rate $E[callback|black = 0]$is 0.097.
Slope: A difference of -0.032 in average callback rates between "white" sounding names and "black sounding names", ie $E[callback|black = 1] - E[callback|black = 0]$.
$E[callback|black = 0] = P[callback|black = 0]$:Callback rate for white sounding names = (When variable "black" is 0) = Intercept of the linear model = 9.7%
$E[callback|black = 1] = P[callback|black = 1]$Callback rate for black sounding names = (When variable "black" is 1) = 0.097-0.032 = 6.5%
There is a 33% decrease in the average callback rate for black sounding names. This may suggest a preference for white sounding names. However, this could depend on various factors like the number of applicants, other factors like sex of the candidate, skillset, years of experience etc.

```
data = read_dta("/Users/anuram/Library/Mobile Documents/com~apple~CloudDocs/MS Stats/Winter 2023/STAT50

ols = lm(data$call ~ data$black)
summary(ols)
```

```
##
## Call:
## lm(formula = data$call ~ data$black)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.09651 -0.09651 -0.06448 -0.06448  0.93552
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.096509   0.005505  17.532  < 2e-16 ***
## data$black  -0.032033   0.007785  -4.115 3.94e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2716 on 4868 degrees of freedom
## Multiple R-squared:  0.003466,   Adjusted R-squared:  0.003261
## F-statistic: 16.93 on 1 and 4868 DF,  p-value: 3.941e-05
```

7(b)

Linear regression can be performed when both X and Y are binary variables, since CEF is linear. X and Y are binary random variable. $E[Y|X=1] = \sum_y yP(Y=y|X=1) = P(Y=1|X=1)$
$E[Y|X=0] = P(Y=1|X=0)$
$E[Y|X=x] = E[Y|X=0] + (E[Y|X=1] - E[Y|X=0])X$
$E[Y|X=x] = \alpha + \beta X$ where $\alpha = E[Y|X=0], \beta = E[Y|X=1] - E[Y|X=0]$
Hence E[Y|X] is linear on X. $\implies CEF = BLP$.

7(c)

The estimates here cannot be interpreted causally since there could be other factors that indicate differences in callback rate for "white or black" sounding names. For example, years of experience or sex of the candidate might also contribute to the call back rate of the candidate. Maybe candidates with "black" sounding names had lower years of experience or "white" sounding names were mostly male candidates. There could be confounding due to these variables and our model does not block for these variables. Hence we cannot interpret the coefficients causally.
Moreover, In regression, we are only predicting the average call back rate given the person's name is black or white sounding. To infer causation, we to set the name of candidate to either white or black sounding names and then predict their call back rate.