Ve406 Lecture 5

Jing Liu

UM-SJTU Joint Institute

May 28, 2018

- Therefore, sample estimates, LSE, and MLE are identical.

$$b_0 = \hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \qquad b_1 = \hat{\beta}_1 = \frac{c_{xy}}{s_x^2}$$

- It is not difficult to show that

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

is a consistent but biased estimator of σ^2, while the following is unbiased

$$\hat{\sigma}^2 = \frac{n}{n-2} s^2$$

though with a larger variance.

- Some author define $\hat{\sigma}^2$ as the sample MSE

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

- Notice the under first set of assumptions, we have

$$\mathbb{E}\left[(Y-\beta_0-\beta_1 X)^2\right]=\sigma^2$$

thus also more or less reaching the same estimator in the same way

$$s^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)^2 \implies \hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)^2$$

Q: Why do we need the stronger set of assumptions? Why MLE?

1. The conditional mean of the response is linear in terms of β_0, β_1, x_i

$$\mathbb{E}\left[Y_i \mid X_i=x_i\right]=\beta_0+\beta_1 x_i$$

2. The errors have zero mean and constant variance

$$\mathbb{E}\left[\varepsilon_i \mid X_i\right]=0 \quad \text{and} \quad \operatorname{Var}\left[\varepsilon_i \mid X_i\right]=\sigma^2 \qquad \text{where} \quad \varepsilon_i=Y_i-\beta_0-\beta_1 X_i$$

3. The errors are independent of X_i, and of each other.
4. The errors follow the normal distribution of $\mathrm{N}\left(0,\sigma^2\right)$.

- Nothing beyond point estimation is possible without sampling distribution of

$$\hat{\beta}_0, \quad \hat{\beta}_1, \quad \text{or} \quad \hat{\sigma}^2$$

- For example, recall the following relationship

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)\, e_i}{\sum_{i=1}^{n} (x_i - \bar{x}_n)^2}$$

Q: Why does $\hat{\beta}_1$ has a normal conditional sampling distribution?

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}\right)$$

- This illustrates that our inference is only as good as our assumptions.

- Most of the followings are questionable if the assumptions are violated.

```
> summary(course.lm)
```

```
Call:
lm(formula = Exam ~ Midterm, data = course.df)

Residuals:
    Min      1Q  Median      3Q     Max
-39.980  -6.471   0.826   8.575  33.242

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)   9.0845     3.2204   2.821  0.00547 **
Midterm       3.7859     0.2647  14.301  < 2e-16 ***
---
Signif. codes:  0 ‘***‘ 0.001 ‘**‘ 0.01 ‘*‘ 0.05 ‘.‘ 0.1 ‘ ‘ 1

Residual standard error: 12.05 on 144 degrees of freedom
Multiple R-squared:  0.5868,    Adjusted R-squared:  0.5839
F-statistic: 204.5 on 1 and 144 DF,  p-value: < 2.2e-16
```

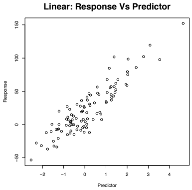
```
> predict(course.lm, data.frame(Midterm = 18),
+              interval = "predict")
```

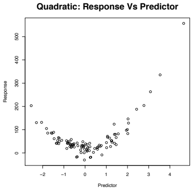
```
       fit      lwr      upr
1 77.23109 53.10062 101.3616
```

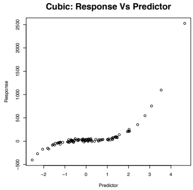
Q: How can we check the first assumption?

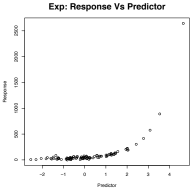
1. The conditional mean of the response is linear in terms of β_0, β_1, x_i

$$\mathbb{E}\left[Y_i \mid X_i = x_i\right] = \beta_0 + \beta_1 x_i$$









- Alternatively, nonlinearity is usually evident in a plot of

$$Y_i \qquad \text{Vs} \qquad \mathbb{E}\left[Y_i \mid X_i = x_i\right]$$

Q: What do you expect to see if the assumption 1. is OK?

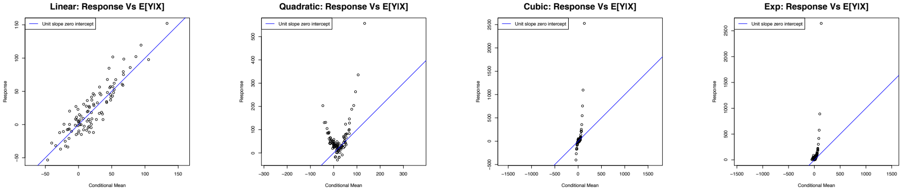
```
> n = 100                                        # Sample size
> beta0 = 17                                     # True intercept
> beta1 = 25                                     # True slope
> df = 10                                        # X parameter
> x.vec = rt(n, df = df)                         # Student t
> s = 10                                         # Y parameter

> ml = beta0 + beta1 * x.vec
> y.ml.vec = rlogis(n, location = ml, scale = s)

> mq = beta0 + beta1 * x.vec^2
> y.mq.vec = rlogis(n, location = mq, scale = s)
>
> mc = beta0 + beta1 * x.vec^3
> y.mc.vec = rlogis(n, location = mc, scale = s)
>
> me = beta0 + beta1 * exp(x.vec)
> y.me.vec = rlogis(n, location = me, scale = s)
```

```
> # Names of 4 different true models
> case.vec = c("Linear", "Quadratic",
+                "Cubic", "Exp")

> # Response for each model
> y.df = data.frame(y.ml.vec, y.mq.vec,
+                   y.mc.vec, y.me.vec)

> pdf() # Plotting Y Vs X, for plots two pages ago
> for (i in 1:ncol(y.df)){
+   tname = bquote(bold(
+     .(case.vec[i])~": Response Vs Predictor"))
+   plot(x.vec, y.df[,i],
+        xlab = "Predictor", ylab = "Response",
+        main = tname, cex.main = 1.8)
+ }
> dev.off()
```

```
null device
          1
```

```r
> pdf() # Plotting Y Vs E[Y|X], for plots next page
> for (i in 1:ncol(y.df)){
+   tname = bquote(bold(
+     .(case.vec[i])~": Response Vs E[Y|X]"))
+
+   plot(ml, y.df[,i],
+        xlab = "Conditional Mean",
+        ylab = "Response",
+        main = tname, cex.main = 1.8, asp = 1)
+
+   abline(a = 0, b = 1, col = "blue")
+
+   legend("topleft", lty = 1, col = "blue",
+          legend = "Unit slope zero intercept")
+ }
> dev.off()
```

```
RStudioGD
        2
```



- Of course, the conditional mean

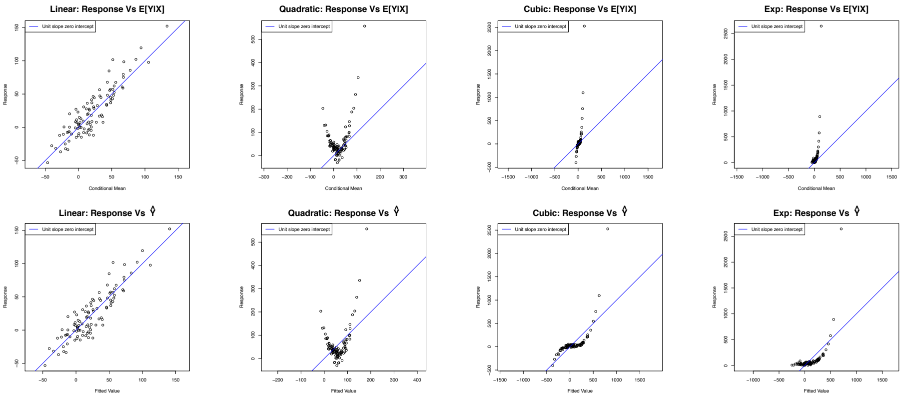
$$\mathbb{E}\left[Y_i \mid X_i = x_i\right] = \beta_0 + \beta_1 x_i$$

is not available outside simulations, we use the estimate of it instead

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

which is known as the fitted value or the predicted value.

- But we again expect points to be scattered around the diagonal line instead of forming a pattern that is systematically off the diagonal line if 1. is true.



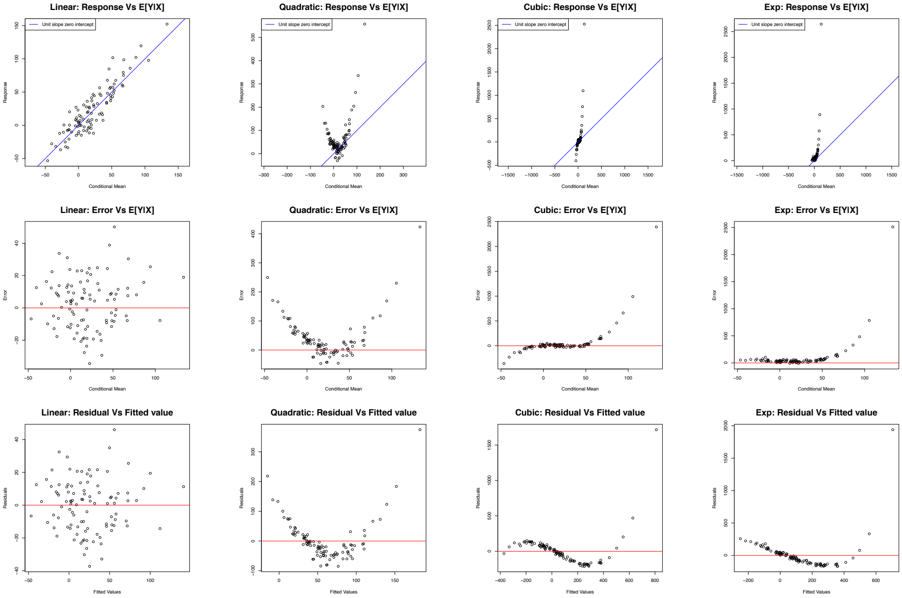
- To avoid the visual distraction of a sloping pattern, we need to consider

$$e_i = y_i - \beta_0 - \beta_1 x_i \qquad \text{Vs} \qquad \beta_0 + \beta_1 x_i$$

In practice, we have to use the estimate of e_i, which is known as the residual

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \qquad \text{Vs} \qquad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

```
> # Plotting Y Vs Yhat, for plots on the second row
> lmlist = list(ml, mq, mc, me)
> pdf()
> for (i in 1:ncol(y.df)){
+   lmlist[[i]] = lm(y.df[,i]~x.vec)
+
+   tname = bquote(bold(
+     .(case.vec[i])~": Response Vs "~hat(Y)))
+
+   plot(lmlist[[i]]$fitted.values, y.df[,i],
+        asp = 1, cex.main = 1.8,
+        xlab = "Fitted Value",
+        ylab = "Response",
+        main = tname)
+   abline(a = 0, b = 1, col = "blue")
+   legend("topleft", lty = 1, col = "Blue",
+          legend = "Unit slope zero intercept")
+ }
> dev.off()
```

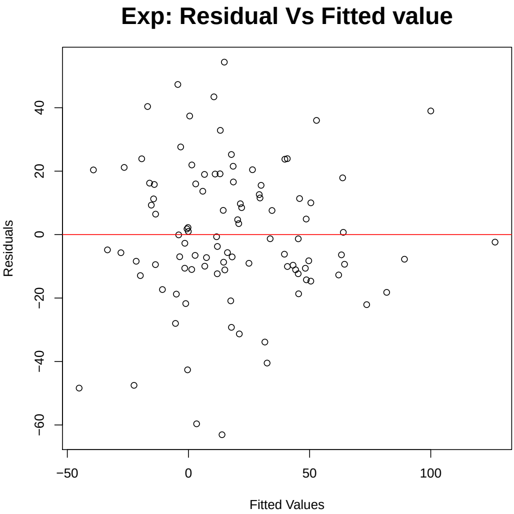

Linear: Response Vs E[Y|X]
Quadratic: Response Vs E[Y|X]
Cubic: Response Vs E[Y|X]
Exp: Response Vs E[Y|X]
Unit slope zero intercept
Response
Conditional Mean
Linear: Error Vs E[Y|X]
Quadratic: Error Vs E[Y|X]
Cubic: Error Vs E[Y|X]
Exp: Error Vs E[Y|X]
Error
Conditional Mean
Linear: Residual Vs Fitted value
Quadratic: Residual Vs Fitted value
Cubic: Residual Vs Fitted value
Exp: Residual Vs Fitted value
Residuals
Fitted Values

```
> pdf() # For plots on the 2nd and 3rd row
> for (i in 1:ncol(y.df)){
+   tname = bquote(bold(
+     .(case.vec[i])~": Error Vs E[Y|X]"))
+   plot(ml, y.df[,i] - ml, cex.main = 1.8,
+         xlab = "Conditional Mean",
+         ylab = "Error", main = tname)
+   abline(h = 0, col = "red")
+
+   tname =
+     bquote(bold(.(case.vec[i])
+                  ~": Residual Vs Fitted value"))
+   plot(lmlist[[i]]$fitted.values,
+         lmlist[[i]]$residuals, cex.main = 1.8,
+         xlab = "Fitted Values",
+         ylab = "Residuals", main = tname)
+   abline(h = 0, col = "red")
+ }
> dev.off()
```

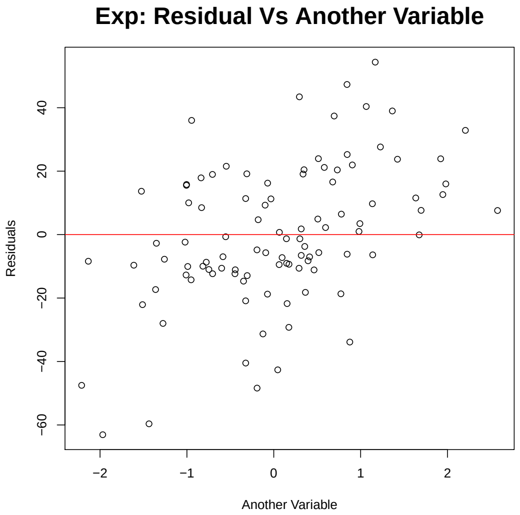
- In practice, you may have other variables in addition to the original predictor

```
> z.vec = rnorm(n)                          # Another variable
> beta2 = 10                                # True parameter
> mmultiple = beta0 + beta1 * x.vec + beta2 * z.vec

> y.mm.vec =
+   rlogis(n, location = mmultiple, scale = s)

> mm.lm = lm(y.mm.vec~x.vec)               # Missing variable

> plot(mm.lm$fitted.values, mm.lm$residuals,
+      cex.main = 1.8,
+      xlab = "Fitted Values",
+      ylab = "Residuals",
+      main = "Exp: Residual Vs Fitted value")
>
> abline(h = 0, col = "red")
```

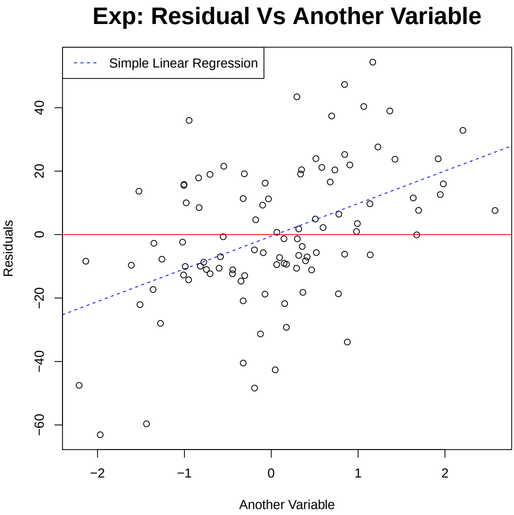
- Notice we don't see any definitively unusual pattern in this case



- However, plotting the residuals against another predictor variable, we see



- Having non-flat band of points is an excellent sign that 1. is false.



```r
> plot(z.vec, mm.lm$residuals,
+       cex.main = 1.8,
+       xlab = "Another Variable",
+       ylab = "Residuals",
+       main = "Exp: Residual Vs Another Variable")
> abline(h = 0, col = "red")
```

```r
> plot(z.vec, mm.lm$residuals,
+       cex.main = 1.8,
+       xlab = "Another Variable",
+       ylab = "Residuals",
+       main = "Exp: Residual Vs Another Variable")
> abline(h = 0, col = "red")
> res.lm = lm(mm.lm$residuals~z.vec)
> abline(res.lm, col = "blue", lty = 2)
> legend("topleft",
+         legend = "Simple Linear Regression",
+         lty = 2, col = "blue")
```

Q: How can we check the second assumption?

2. The errors have zero mean and constant variance

$$\mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0 \quad \text{and} \quad \mathrm{Var}\left[\varepsilon_i \mid X_i\right] = \sigma^2 \qquad \text{where} \quad \varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$$

- Since the errors e_i are not directly observed, we consider the residual

$$\begin{aligned}\hat{e}_i &= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= \beta_0 + \beta_1 x_i + e_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ &= \left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right) x_i + e_i\end{aligned}$$

- The terms in the parentheses are hopefully small, but they are not usually 0.
- Recall we have the following

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}_n}{(n-1)s_x^2}\right) e_i \quad \text{and} \quad \hat{\beta}_0 = \beta_0 + \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{(n-1)s_x^2}\right) e_i$$

- Putting everything together, we have

$$\hat{e}_i = \sum_{j=1}^{n} \left(\frac{1}{n} - \frac{\bar{x}(x_j - \bar{x})}{(n-1)s_x^2} \right) e_j + x_i \sum_{j=1}^{n} \left(\frac{x_j - \bar{x}_n}{(n-1)s_x^2} \right) e_j + e_i = \sum_{j=1}^{n} c_{ij} e_j$$

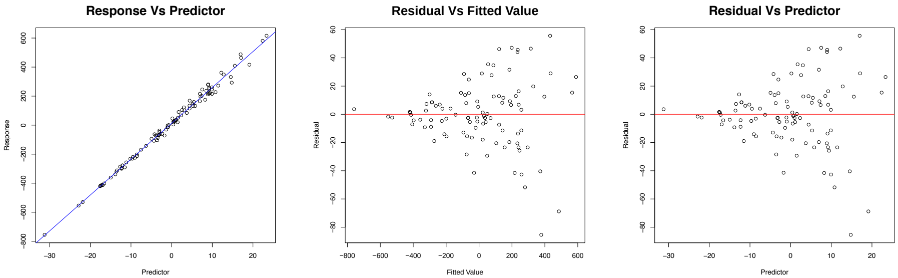
where c_{ij} depends only on n and x_1, x_2, ..., x_n.

Q: So, we have

$$\mathbb{E}\left[\hat{e}_i \mid X\right] = \sum_{j=1}^{n} c_{ij} \mathbb{E}\left[e_j \mid X\right] = 0$$

$$\operatorname{Var}\left[\hat{e}_i \mid X\right] = \sum_{j=1}^{n} c_{ij}^2 \operatorname{Var}\left[e_j \mid X\right] = \sigma^2 \sum_{j=1}^{n} c_{ij}^2$$

thus we expect points in residual Vs predictor, or residual Vs fitted value, plot to be scattered around x-axis within roughly the same bandwidth.



- Notice the second plot and third are essentially the same in this case.
- It is clear that the constant variance assumption is not satisfied.

```
> x.new.vec = rnorm(n, mean = 0, sd = 10)
> msl = beta0 + beta1 * x.new.vec
> y.msl.vec = rnorm(n, mean = msl,
+                      sd = abs(x.new.vec+20))
> msl.lm = lm(y.msl.vec~x.new.vec)
```

Q: How can we check the third assumption?

3. The errors are independent of X_i, and of each other.

- You might be tempted to use the fact the following must be zero if 3. is true

$$\mathrm{Cov}\left[X_i, \varepsilon_i\right] = 0$$

and thus the sample covariance shall not be too far away from zero, that is,

$$\sum_{i=1}^{n} \left(e_i - \bar{e}_n\right)\left(x_i - \bar{x}_n\right) = \sum_{i=1}^{n} e_i \left(x_i - \bar{x}_n\right) \approx 0$$

which is correct, however, you might wrongly jump to the conclusion to test

$$\sum_{i=1}^{n} \left(\hat{e}_i - \bar{\hat{e}}_n\right)\left(x_i - \bar{x}_n\right) = \sum_{i=1}^{n} \hat{e}_i \left(x_i - \bar{x}_n\right)$$

Q: Why is this not going to provide any information about $\mathrm{Cov}\left[X_i, \varepsilon_i\right]$?

- Recall $\hat{\beta}_0$ and $\hat{\beta}_1$ are solutions to the following estimation equations

$$\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0 \implies \sum_{i=1}^{n}\left(\hat{\beta}_0+\hat{\beta}_1 x_i+\hat{e}_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0$$

$$\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)(x_i)=0 \implies \sum_{i=1}^{n}\hat{e}_i x_i=0$$

- Hence we see the sample covariance between residual and X is always 0

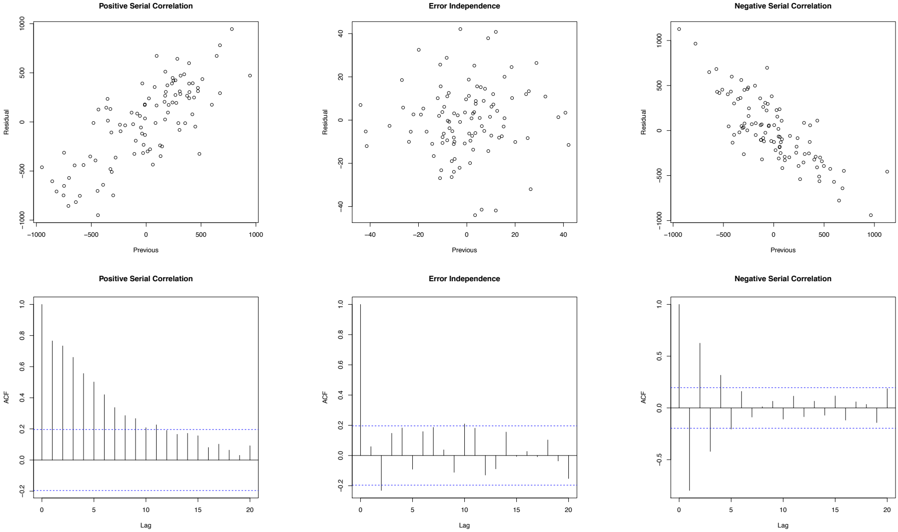
$$\sum_{i=1}^{n}\hat{e}_i=0 \implies \sum_{i=1}^{n}\hat{e}_i x_i-\bar{x}_n\sum_{i=1}^{n}\hat{e}_i=0 \implies \sum_{i=1}^{n}\hat{e}_i\left(x_i-\bar{x}_n\right)=0$$

- Therefore, other than we have done for 1., the independence between X and ε cannot be further checked using any diagnostics.

Q: How about the part says that errors are independent?

$$\sum_{i=1}^{n}\hat{e}_i=0$$

- We expect small nonzero sample covariances amongst residuals, but we do not expect to see strong relation amongst residuals if 3. is true.



```
> # Generating autocorrelated data
> y.psc.vec = double(n)
> y.nsc.vec = double(n)
>
> y.psc.vec[1] = rnorm(1, mean = msl[1])
> y.nsc.vec[1] = rnorm(1, mean = msl[1])
>
> for (i in 2:n) {
+   y.psc.vec[i] =
+     rnorm(1, mean = msl[i] + 0.8 * y.psc.vec[i-1])
>
+   y.nsc.vec[i] =
+     rnorm(1, mean = msl[i] - 0.8 * y.nsc.vec[i-1])
+ }
>
> psc.lm = lm(y.psc.vec~x.new.vec)
> nsc.lm = lm(y.nsc.vec~x.new.vec)
```

```
# Plotting ehat_i against ehat_{i-1}, and acf
> plot(psc.lm$residuals[-n],psc.lm$residuals[-1],
+      xlab = "Previous", ylab = "Residual",
+      main = "Positive Serial Correlation")
> plot(lmlist[[1]]$residuals[-n],
+      lmlist[[1]]$residuals[-1],
+      xlab = "Previous", ylab = "Residual",
+      main = "Error Independence")
> plot(nsc.lm$residuals[-n],nsc.lm$residuals[-1],
+      xlab = "Previous", ylab = "Residual",
+      main = "Negative Serial Correlation")
>
> acf(psc.lm$residuals,
+     main = "Positive Serial Correlation")
> acf(lmlist[[1]]$residuals,
+     main = "Error Independence")
> acf(nsc.lm$residuals,
+     main = "Negative Serial Correlation")
```