# STAT 3008 Applied Regression Analysis
# Tutorial 3.

XU yongze

Feb 23 & 24 & 26, 2013

## Contents

# 1 Review Questions

For model $y_i = \beta_0 + \beta_1 x_i + e_i, e_i \sim i.i.d.N(0, \sigma^2)$, derive the followings:

a. $\sum y_i = \sum \hat{y}_i$

b. $\sum x_i \hat{e}_i = 0$

c. $\sum \hat{y}_i \hat{e}_i = 0$

d. $\rho(\hat{\beta}_0, \hat{\beta}_1) = \frac{\bar{x}}{\sqrt{\frac{SXX}{n} + \bar{x}^2}}$.

Hint: In matrix notation,

$$\mathbf{y} = X\beta + \mathbf{e},$$

$H = X(X'X)^{-1}X'$ is symmetric and idempotent,

$$\hat{\mathbf{y}} = H\mathbf{y},$$

For a)

$$\sum y_i = \sum \hat{y}_i \Leftrightarrow \mathbf{1}'\mathbf{y} = \mathbf{1}'H\mathbf{y}, \mathbf{1}' = (1 \cdots 1),$$

$$HX = X, X = (\mathbf{1} \ *), H\mathbf{1} = \mathbf{1}.$$

For b)

$$X'\hat{\mathbf{e}} = \mathbf{0} \Rightarrow \sum x_i \hat{e}_i = 0,$$

$$X'\hat{\mathbf{e}} = X'(I - H)\mathbf{y} = (X' - X')\mathbf{y} = \mathbf{0}.$$

For c)

$$\hat{\mathbf{y}}'\hat{\mathbf{e}} = (H\mathbf{y})'(I - H)\mathbf{y} = \mathbf{y}'(H - H)\mathbf{y} = 0$$

Not in matrix notation,

a)

$$\begin{aligned}
&\sum \hat{y}_i \\
&= \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= \sum ((\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 x_i) \\
&= n\bar{y} - n\hat{\beta}_1 \bar{x} + \sum \hat{\beta}_1 x_i = n\bar{y} \\
&= \sum y_i
\end{aligned}$$

b)

$$
\begin{aligned}
&\sum x_i \hat{e_i} \\
&= \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \\
&= \sum x_i y_i - \hat{\beta}_0 \sum x_i - \hat{\beta}_1 \sum x_i^2 \\
&= \sum x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x})(n\bar{x}) - \hat{\beta}_1 \sum x_i^2 \\
&= \sum x_i y_i - n\bar{x}\bar{y} + \hat{\beta}_1 (n\bar{x}^2 - \sum x_i^2) \\
&= \sum x_i y_i - n\bar{x}\bar{y} - \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} (\sum x_i^2 - n\bar{x}^2) \\
&= 0
\end{aligned}
$$

c)

$$
\begin{aligned}
&\sum \hat{y}_i \hat{e_i} \\
&= \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i) \hat{e_i} \\
&= \hat{\beta}_0 \sum \hat{e_i} + \hat{\beta}_1 \sum x_i \hat{e_i} \\
&= 0
\end{aligned}
$$

d)

$$
\begin{aligned}
&\rho(\hat{\beta}_0, \hat{\beta}_1) \\
&= \frac{Cov(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{Var(\hat{\beta}_0) Var(\hat{\beta}_1)}} \\
&= \frac{-\frac{\bar{x}}{SXX}\sigma^2}{\sqrt{(\frac{1}{n} + \frac{\bar{x}^2}{SXX})\sigma^2 (\frac{1}{SXX})\sigma^2}} \\
&= -\frac{\bar{x}}{SXX} * \frac{1}{\sqrt{(\frac{SXX + n\bar{x}^2}{nSXX}) \frac{1}{SXX}}} \\
&= \frac{\bar{x}}{\sqrt{\frac{SXX}{n} + \bar{x}^2}}
\end{aligned}
$$

# 2 ANOVA Test(Analysis of Variance)

♣ What is ANOVA test?

A test to compare different models, such as E(Y|X) = $\beta_0$ and E(Y|X) = $\beta_0 + \beta_1 x$.

♣ Comparing Models.

$H_0 : E(Y|X) = \beta_0 \; v.s. \; H_1 : E(Y|X) = \beta_0 + \beta_1 x$

Under $H_0$, $\beta_1 = 0$, which means x and y are independent.

We have two methods to finish the comparison:

(1) Test $\beta_1 = 0$

(2) ANOVA Test:

Under $0, \hat{\beta}_0 = \bar{y}, RSS_0 = \sum (y_i - \bar{y})^2 = SYY.$

Under $1, RSS_1 = \sum [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x)]^2 = SYY - \frac{SXY^2}{SXX}$

Obviously, $RSS_0 > RSS_1$, but it is not necessarily model under $H_1$ is better.

$SS_{reg}$ = Difference sum of squares due to regression.(as we add another parameter $\beta_1$ to the model).

$SS_{reg} = \triangle RSS = RSS_0 - RSS_1 = \frac{SXY^2}{SXX}$

If $SS_{reg}$ is large, which means the addition of the parameter $\beta_1$ is useful to explain the variation of the data.

Q:How large is large?

Notice:

$$\frac{SS_{reg}}{\sigma^2} \sim \chi_1^2 \tag{2.1}$$

$$\frac{RSS_1}{\sigma^2} \sim \chi_{n-2}^2 \tag{2.2}$$

We can derive:

$$T.S. = \frac{SS_{reg}}{\frac{RSS_1}{n-2}} \sim F_{1,n-2}. \tag{2.3}$$

If obs.T.S. > $F_{(\alpha,1,n-2)}$, $H_0$ is rejected at $\alpha$ level.

The following is the very important ANOVA Table.

| Source | DF | SS | MS | F |
|---|---|---|---|---|
| Regression | 1 | $SS_{reg} = \triangle RSS = \frac{SXY^2}{SXX}$ | $MS_{reg} = \frac{SS_{reg}}{1} = SS_{reg}$ | $\frac{MSR}{MSE}$ |
| Residuals(Error) | n-2 | $SSE = RSS_1 = SYY - \frac{SXY^2}{SXX}$ | MSE = $\frac{SSE}{n-2}$ | |
| Total | n-1 | $SST = RSS_0 = SYY$ | | |

# 3 Hypothesis Testing of Parameter Estimates

1)$\beta_0$**(Intercept)**:

$H_0 : \beta_0 = \beta_0^* \; v.s. \; H_1 : \beta_0 \neq \beta_0^*$

$$T.S. = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\hat{\sigma^2}(\frac{1}{n} + \frac{\bar{x}^2}{SXX})}} \sim t_{n-2} \tag{3.1}$$

If|obs.T.S.| > $t_{(\frac{\alpha}{2}, n-2)}$, $H_0$ is rejected at $\alpha$ level.

2)$\beta_1$**(Slope):**

$H_0 : \beta_1 = \beta_1^* \; v.s. \; H_1 : \beta_1 \neq \beta_1^*$

$$T.S. = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\hat{\sigma^2}\frac{1}{SXX}}} \sim t_{n-2} \tag{3.2}$$

If|obs.T.S.| > $t_{(\frac{\alpha}{2}, n-2)}$, $H_0$ is rejected at $\alpha$ level.

3)$\hat{y}_*|x_*$**(Prediction):**

$H_0 : y_*|x_* = y_* \; v.s. \; H_1 : y_*|x_* \neq y_*$

$$T.S. = \frac{\hat{y}_*|x_* - y_*}{\sqrt{\hat{\sigma^2}(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})}} \sim t_{n-2} \tag{3.3}$$

If|obs.T.S.| > $t_{(\frac{\alpha}{2}, n-2)}$, $H_0$ is rejected at $\alpha$ level.

4)$\hat{y}|x_*$**(Fitted Value):** (Just for your reference)

$H_0 : E(y|x_*) = y_* \; v.s. \; H_1 : E(y|x_*) \neq y_*$

$$T.S. = \frac{(\hat{y}|x_* - y_*)^2}{2\hat{\sigma^2}(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})} \sim F_{(2,n-2)} \tag{3.4}$$

If obs.T.S. > $F_{(\alpha,2,n-2)}$, $H_0$ is rejected at $\alpha$ level.

5) $\sigma^2$**(Variance):**

$H_0 : \sigma^2 = \alpha \; v.s. \; H_1 : \sigma^2 > \alpha$

$$T.S. = \frac{(n-2)\hat{\sigma^2}}{a} \sim \chi^2_{n-2} \tag{3.5}$$

If obs.T.S. > $\chi^2_{(\alpha,n-2)}$, $H_0$ is rejected at $\alpha$ level.

# 4 Confidence Interval of Parameter Estimates

Construction of Confidence Interval

1)$\beta_0$**(Intercept):**

$$\hat{\beta}_0 \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{SXX})) \tag{4.1}$$

$$(n-2)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2} \tag{4.2}$$

We can derive:

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\widehat{\sigma^2}(\frac{1}{n} + \frac{\bar{x}^2}{SXX})}} \sim t_{n-2} \tag{4.3}$$

Then, the (1-$\alpha$)100%confidence interval of $\beta_0$ is:

$$\left[\hat{\beta}_0 - t_{(n-2,\frac{\alpha}{2})}\sqrt{\widehat{\sigma^2}(\frac{1}{n} + \frac{\bar{x}^2}{SXX})},\ \hat{\beta}_0 + t_{(n-2,\frac{\alpha}{2})}\sqrt{\widehat{\sigma^2}(\frac{1}{n} + \frac{\bar{x}^2}{SXX})}\right] \tag{4.4}$$

2)$\beta_1$**(Slope):**

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2\frac{1}{SXX}) \tag{4.5}$$

$$(n-2)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2} \tag{4.6}$$

We can derive:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{\sigma^2}\frac{1}{SXX}}} \sim t_{n-2} \tag{4.7}$$

Then, the (1-$\alpha$)100%confidence interval of $\beta_1$ is:

$$\left[\widehat{\beta_1} - t_{(n-2,\frac{\alpha}{2})}\sqrt{\widehat{\sigma^2}\frac{1}{SXX}},\ \widehat{\beta_1} + t_{(n-2,\frac{\alpha}{2})}\sqrt{\widehat{\sigma^2}\frac{1}{SXX}}\right] \tag{4.8}$$

3)$\hat{y}_* \mid x_*$ **(Prediction):**

Prediction $\hat{y}_*|x_*$ is the value of y given that value of $x_*$ is **observed**.

$$\hat{y_*}|x_* = \hat{\beta}_0 + \hat{\beta}_1 x_* \tag{4.9}$$

Thus, variance of prediction is the deviation of prediction value and the mean value of y.

$$Var(\hat{y_*}|x_*) = \sigma^2(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}) \tag{4.10}$$

$$\widehat{Var(\hat{y_*}|x_*)} = \hat{\sigma^2}(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}) \tag{4.11}$$

Therefore, the prediction interval (confidence interval) of $\hat{y_*}|x_*$ is:

$$\left[\hat{y_*}|x_* - t_{(n-2,\frac{\alpha}{2})}\sqrt{\hat{\sigma^2}(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})},\ \hat{y_*}|x_* + t_{(n-2,\frac{\alpha}{2})}\sqrt{\hat{\sigma^2}(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})}\right] \tag{4.12}$$

4)$\hat{y}|x_*$**(Fitted Value):**

Fitted value is:

$$E(Y|X = x_*) = \hat{\beta}_0 + \hat{\beta}_1 x_* \tag{4.13}$$

Here we estimate the **mean of the Y population** associated with $x_*$.

$$Var(\hat{y}|x_*) = \sigma^2(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}) \tag{4.14}$$

$$\widehat{Var(\hat{y}|x_*)} = \hat{\sigma^2}(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX}) \tag{4.15}$$

The confidence interval of the fitted value is:

$$\hat{y}|x_* \pm t_{(n-2,\frac{\alpha}{2})}\sqrt{\hat{\sigma^2}(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})} \tag{4.16}$$

The simultaneous confidence band of $\hat{y}|x_*$ is:

$$\left[\hat{y}|x_* - [2F_{(a,2,n-2)}]^{\frac{1}{2}}\sqrt{\widehat{\sigma^2}(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})},\ \hat{y}|x_* + [2F_{(a,2,n-2)}]^{\frac{1}{2}}\sqrt{\widehat{\sigma^2}(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})}\right] \tag{4.17}$$

(Please refer to "Statistical Inference" by G. Casella & R. Berger ).

5)$\sigma^2$**(Variance):**

$$(n-2)\frac{\widehat{\sigma^2}}{\sigma^2} \sim \chi^2_{n-2} \tag{4.18}$$

The (1-$\alpha$)100%confidence interval of $\sigma^2$ is:

$$[\frac{(n-2)\hat{\sigma^2}}{\chi^2_{(1-\frac{\alpha}{2},n-2)}}, \frac{(n-2)\hat{\sigma^2}}{\chi^2_{(\frac{\alpha}{2},n-2)}}] \tag{4.19}$$

# 5 Exercises

## 5.1 Exercise 3 Q1

## 5.2 Exercise 3 Q2

## 5.3 Example from 12-13 Midterm(similar to Ex3 Q3)

In simple linear regression, if the value of the predictor $X$ is replaced by $cX$, where $c$ is some non-zero constant, which of the following will be affected? (Circle the answer(s))
(a)$\hat{\beta}_0$, (b)$\hat{\beta}_1$, (c)$\hat{\sigma}^2$, (d)$R^2$, (e)t-test statistic of the null hypothesis $H_0 : \beta_1 = 0$.
Hint:

$$\hat{\beta}_{1_{(new)}} = \frac{c\sum(x_i-\bar{x})(y_i-\bar{y})}{c^2\sum(x_i-\bar{x})^2} = \frac{1}{c}\hat{\beta}_1.$$

$$\hat{\beta}_{0_{(new)}} = \bar{y} - (\frac{1}{c}\hat{\beta}_1)(c\bar{x}) = \bar{y} - \hat{\beta}_1\bar{x} = \hat{\beta}_0.$$

$$\hat{e}_{i_{(new)}} = y_i - \hat{\beta}_0 - \hat{\beta}_{1_{(new)}}x_{i_{(new)}} = y_i - \hat{\beta}_0 - (\frac{1}{c}\hat{\beta}_1)(cx_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = \hat{e}_i,$$

$$\therefore \hat{\sigma}^2_{(new)} = \frac{\sum \hat{e}_i^2}{n-2} = \hat{\sigma}^2.$$

$$\because \hat{e}_{i_{(new)}} = \hat{e}_i, RSS_{(new)} = RSS.$$

$$\because SYY = \sum(y_i-\bar{y})^2 \text{ is unaffected,}$$

$$\therefore R^2_{(new)} = 1 - \frac{RSS}{SYY} = R^2.$$

$$t_{(new)} = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)_{(new)}}, \text{ where } se(\hat{\beta}_0)_{(new)} = \hat{\sigma}(\frac{1}{n} + \frac{\bar{x}^2_{(new)}}{SXX_{(new)}})^{1/2}.$$

$$\because \frac{\bar{x}^2_{(new)}}{SXX_{(new)}} = \frac{c^2\bar{x}^2}{c^2\sum(x_i-\bar{x})} = \frac{\bar{x}^2}{\sum(x_i-\bar{x})} = \frac{\bar{x}^2}{SXX},$$

$$\therefore t_{(new)} = t.$$

## 5.4 Exercise 3 Q5 2.12.4

## 5.5 Example from 11-12 Midterm

1. (30) Fill in the missing values in the following tables of regression output .

| ANOVA Table | | | | | |
|---|---|---|---|---|---|
| Source | Sum of Squares | d.f. | Mean Square | F-statistics | p-value |
| Regression | _____ | _____ | _____ | _____ | 9.5e-09 |
| Residuals | _____ | _____ | _____ | | |
| Total | _____ | | | | |

| Coefficient Table | | | | |
|---|---|---|---|---|
| Variable | Coefficient | s.e. | t-statistics | p-value |
| Constant | _____ | _____ | _____ | 0.00322 |
| X | -2.04245 | _____ | _____ | _____ |
| $n$=_____ | $\hat{\sigma}$ = _____ | $R^2$=_____ | | |

p.s. It is found in R that $qf(1-9.5e^{-9}, 1, 6) = 1917.3$. Also $\bar{x} = 5.125$, $\bar{y} = -9.1974$, $SXX = 54.875$.

**Solution:**

ANOVA Table

| Source | SS | Df | MS | F | p-value |
|---|---|---|---|---|---|
| Regression | 228.9167 | 1 | 228.9167 | 1917.3 | 9.5e-09 |
| Residuals | 0.7164 | 6 | 0.1194 | | |
| Total | 229.6331 | | | | |

Coefficient Table

| Variable | Coefficient | s.e. | t | p-value |
|---|---|---|---|---|
| Constant | 1.2702 | 0.2685 | 4.7313 | 0.00322 |
| X | -2.04245 | 0.04665 | 43.7861 | 9.5e-09 |
| $n = 8$ | $\hat{\sigma} = 0.3455$ | $R^2 = 0.9969$ | | |

$\because qf(1 - 9.5 \times 10^{-9}, 1, 6) = 1917.3$

$\therefore F = 1917.3$

$t^2_{\beta_1} = F, \ \therefore |t_{\beta_1}| = \sqrt{1917.3} = 43.7861$

the corresponding $\mathrm{p} - \mathrm{value} = 9.5^{-9}$

$df_{reg} = 1, df_{residuals} = 6, n = df_{residuals} + 2 = 8.$

$s.e.(\hat{\beta}_1) = |\hat{\beta}_1 / |t_{\beta_1}|| = 0.04665$

$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = 1.2702$

$SXY = \hat{\beta}_1 SXX = -112.0794$

$$SS_{reg} = \frac{SXY^2}{SXX} = 228.9167$$

$MS_{reg} = SS_{reg}/1 = 228.9167$

$MS_{residuals} = MS_{reg}/F = 0.1194$

$SS_{residuals} = 6 \times MS_{residuals} = 0.7164$

$SS_{total} = SS_{residuals} + SS_{reg} = 229.6331$

$\hat{\sigma}^2 = MS_{residuals} = 0.1194, \therefore \hat{\sigma} = 0.3455$

$$R^2 = \frac{SS_{reg}}{SS_{total}} = 0.9969$$

$$s.e.(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SXX}} = 0.2685$$

$$t_{\beta_0} = \frac{\hat{\beta}_0}{s.e.(\hat{\beta}_0)} = 4.7313$$