EMSE 6765: DATA ANALYSIS

For Engineers and Scientists

Session 12: Comparing Imbedded Models, Forecasting

Version: 4/05/2021

Lecture Notes by: J. René van Dorp[1]

www.seas.gwu.edu/~dorpjr

---

[1] Department of Engineering Management and Systems Egineering, School of Engineering and Applied Science, The George Washington University, 800 22nd Street, N.W., Suite 2800, Washington D.C. 20052. E-mail: dorpjr@gwu.edu.

Regression Analysis: Log(Price) versus Elevation, Sewer, Date, Flood

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	2.2320	0.55800	21.49	0.000
Error	26	0.6753	0.02597		
Total	30	2.9072			

Model Summary

S	R-sq	R-sq(adj)
0.161156	76.77%	73.20%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	1.4891	0.0915	16.28	0.000
Elevation	0.01411	0.00816	1.73	0.096
Sewer	-0.000044	0.000014	-3.26	0.003
Date	0.00741	0.00122	6.05	0.000
Flood	-0.3183	0.0887	-3.59	0.001

Regression Equation

Log(Price) = 1.4891 + 0.01411 Elevation - 0.000044 Sewer + 0.00741 Date - 0.3183 Flood

- Prediction of Log(Price) using **the smaller model:**

$$Y = \boldsymbol{x}_0^T \widehat{\boldsymbol{b}} + \epsilon,\ E[\epsilon] = 0, \epsilon \sim \boldsymbol{N(0,\sigma)} \Leftrightarrow E[Y|\boldsymbol{x}_0] = \boldsymbol{x}_0^T \widehat{\boldsymbol{b}}$$

Prediction for Log(Price)

Regression Equation

Log(Price) = 1.4891 + 0.01411 Elevation - 0.000044 Sewer + 0.00741 Date - 0.3183 Flood

Settings

Variable	Setting
Elevation	0
Sewer	0
Date	0
Flood	0

Prediction

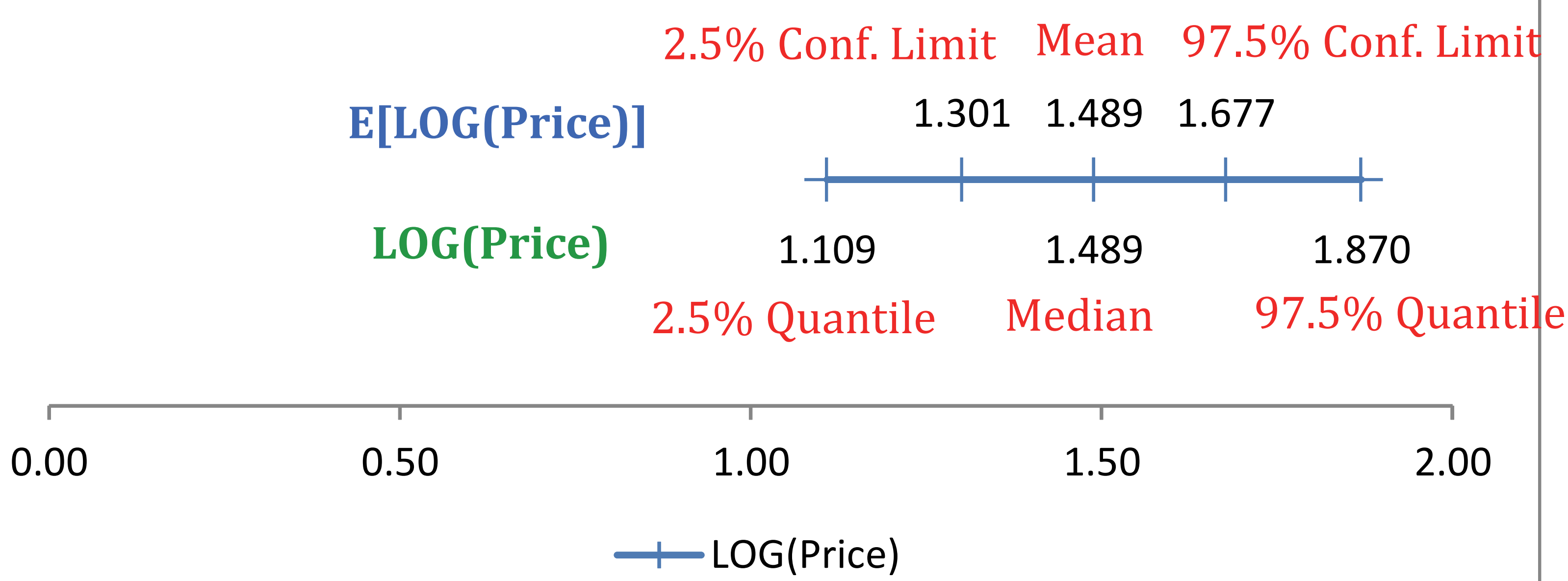
Fit	SE Fit	95% CI	95% PI
1.48907	0.0914849	(1.30102, 1.67712)	(1.10816, 1.86999)

$\widehat{y} = x_0^T \widehat{b} \approx 1.48907$ is **both a prediction for r.v. Y and mean $E[Y|x_0]$**

$(1.301, 1.677)$ is **conf. interval for true mean $E[Y|x_0]$**, no prob. interpretation

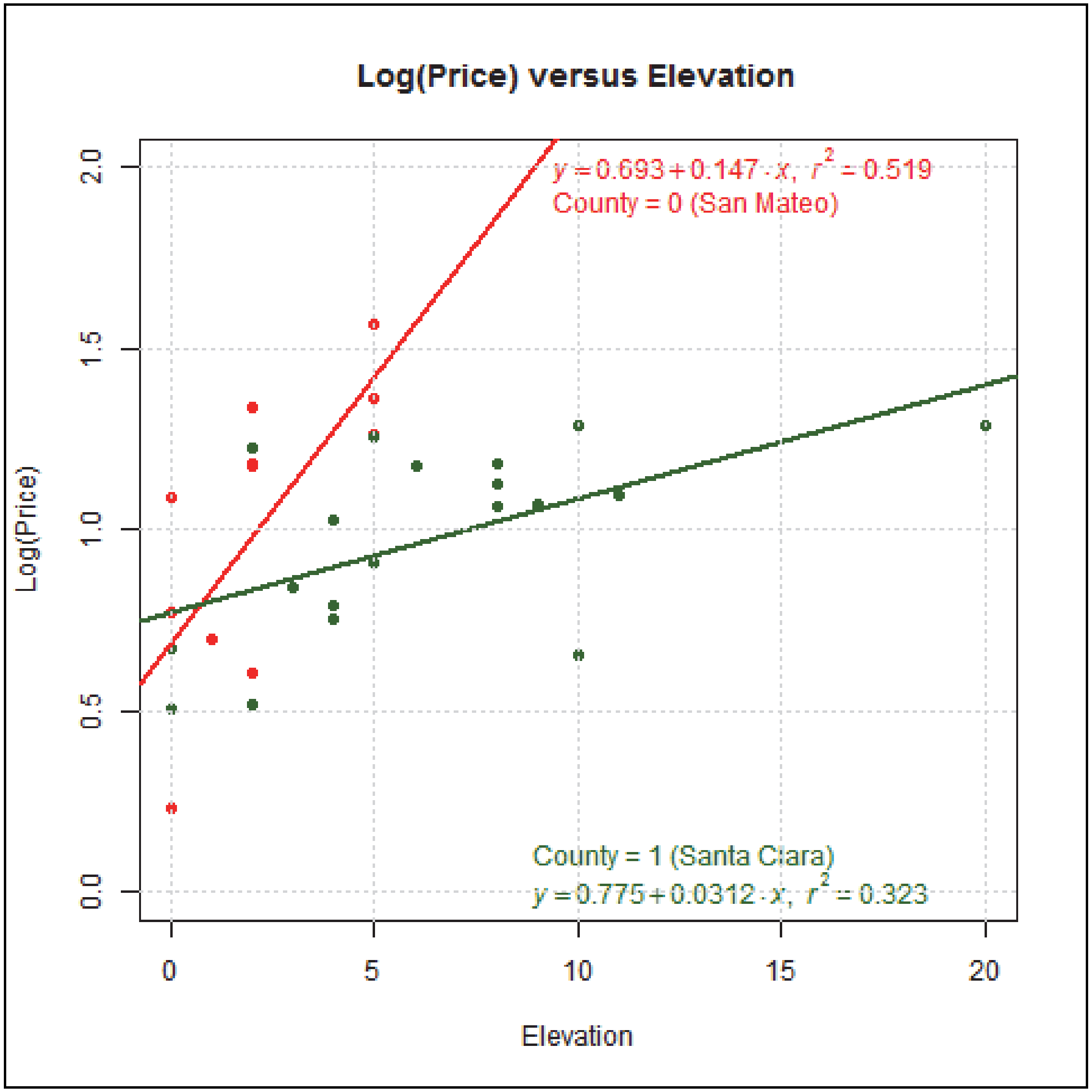
$(1.108, 1.870)$ is **pred./cred. interval for r.v. Y**, with prob. interpretation

Fit	SE Fit	95% CI	95% PI
1.48907	0.0914849	(1.30102, 1.67712)	(1.10816, 1.86999)



Observe the Median and the Mean are of the same value! Why?

$$Y = \boldsymbol{x}_0^T \widehat{\boldsymbol{b}} + \epsilon,\ E[\epsilon] = 0, \epsilon \sim N(0, \sigma) \Leftrightarrow E[Y|\boldsymbol{x}_0] = \boldsymbol{x}_0^T \widehat{\boldsymbol{b}}$$


Log(Price) versus Elevation
y = 0.693 + 0.147 · x, r² = 0.519
County = 0 (San Mateo)
County = 1 (Santa Clara)
y = 0.775 + 0.0312 · x, r² = 0.323
Log(Price)
0.0
0.5
1.0
1.5
2.0
0
5
10
15
20
Elevation

Suggestion: **Capture interaction effect using an interaction term**

$$log(PRICE) = b_0 + b_1 ELEVATION + b_2 SEWER + b_3 DATE + b_4 FLOOD + b_5 COUNTY + b_6(COUNTY \times ELEVATION)$$

When $COUNTY = 0$ **the above equation reduces to**

$$log(PRICE) = \boldsymbol{b_0} + \boldsymbol{b_1} ELEVATION + b_2 SEWER + b_3 DATE + b_4 FLOOD.$$

When $COUNTY = 1$ **the above equation reduces to**

$$log(PRICE) = (\boldsymbol{b_0} + \boldsymbol{b_5}) + (\boldsymbol{b_1} + \boldsymbol{b_6}) ELEVATION + b_2 SEWER + b_3 DATE + b_4 FLOOD$$

Thus, **the interaction effect here allows for different intercepts and slopes by counties**.

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.146820	82.20%	77.76%	69.05%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.424	0.127	11.19	0.000	
Elevation	0.0483	0.0286	1.69	0.104	21.57
Sewer	-0.000048	0.000012	-3.86	0.001	1.32
Date	0.00548	0.00135	4.06	0.000	1.52
Flood	-0.394	0.102	-3.88	0.001	2.00
County	-0.113	0.110	-1.03	0.312	4.11
County*Elevation	-0.0307	0.0297	-1.03	0.312	28.12

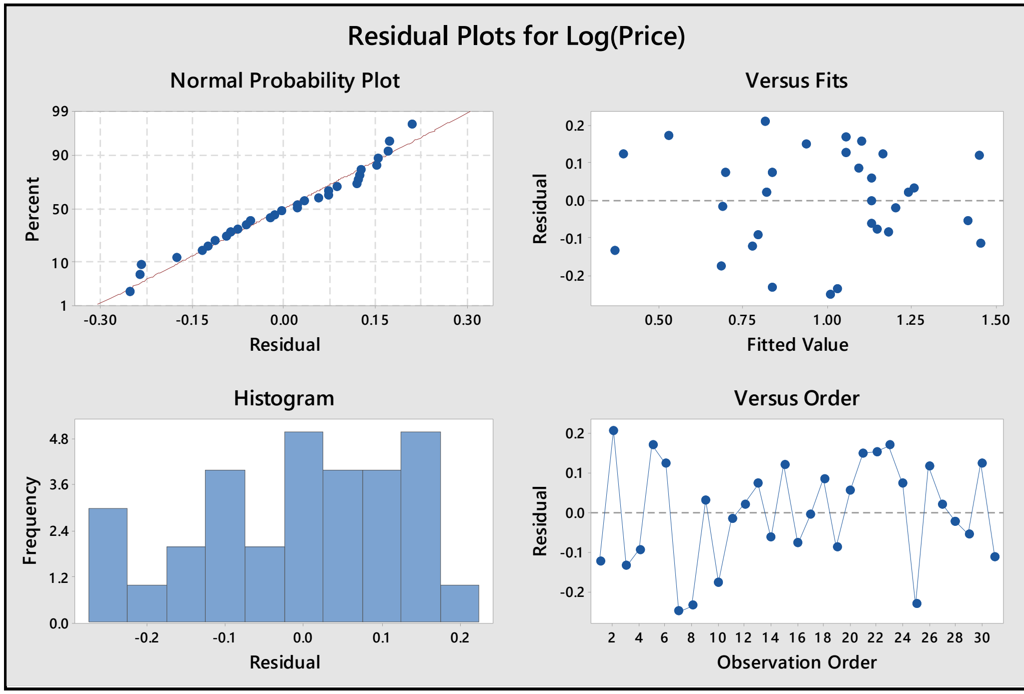
Regression Equation

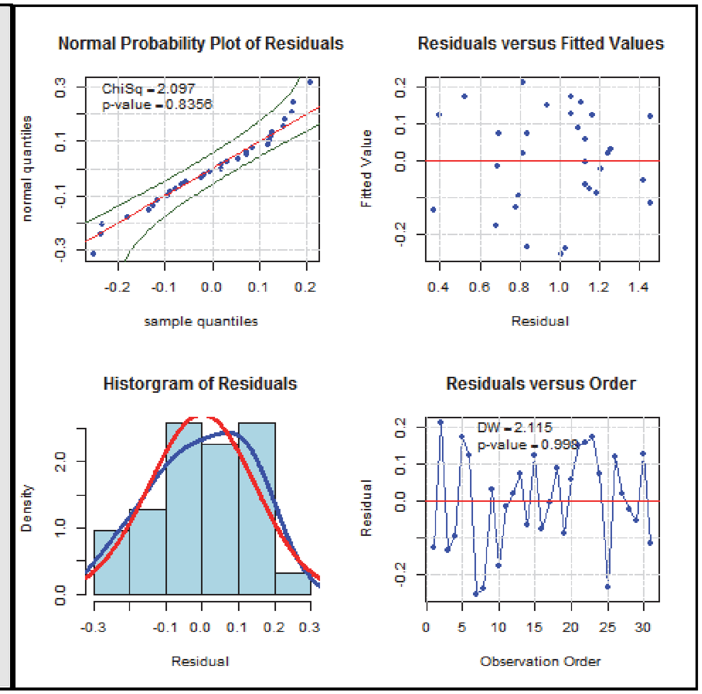
Log(Price) = 1.424 + 0.0483 Elevation - 0.000048 Sewer + 0.00548 Date - 0.394 Flood - 0.113 County - 0.0307 County*Elevation

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.11482

$R^2_{adj} = 77.8\% \uparrow$ (Previously $R^2_{adj} = 73.2\%$) , DW-Statistic ≈ 2.11





Normality and independence assumption of residuals seem reasonable. (although we can observe one outlier)

$$e_2^* \approx 1.72, \mathrm{DFIT}_2 \approx 1.02 > 2\sqrt{7/31} \approx 0.95 \Rightarrow \text{Should be checked.}$$

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.146820	82.20%	77.76%	69.05%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.424	0.127	11.19	0.000	
Elevation	0.0483	0.0286	1.69	0.104	21.57
Sewer	-0.000048	0.000012	-3.86	0.001	1.32
Date	0.00548	0.00135	4.06	0.000	1.52
Flood	-0.394	0.102	-3.88	0.001	2.00
County	-0.113	0.110	-1.03	0.312	4.11
County*Elevation	-0.0307	0.0297	-1.03	0.312	28.12

Regression Equation

Log(Price) = 1.424 + 0.0483 Elevation - 0.000048 Sewer + 0.00548 Date - 0.394 Flood - 0.113 County - 0.0307 County*Elevation

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.11482

Is the improvement in the R^2-values from 76.7% to 82.2% (a jump of about 5.4%) statistically significant ?

- When the simpler model description is **completely contained within** the description of the larger model, we can perform **an F-hypothesis test**:

Explanatory Variables in the full model					
County	Elevation	Sewer	Date	Flood	County*Elevation
Explanatory Variables in the restricted/small model					
	Elevation	Sewer	Date	Flood	
Conclusion: All variables of small/restricted model are variables in the full model					
and "the increase in R^2 test" an be perfromed					

$\boldsymbol{H_0}$ **: No model improvement ,** $\boldsymbol{H_1}$ **: Model Improvement**

R_f^2 : R^2-value of **the full model**, R_r^2 : R^2-value of **restricted model**

df_f : Degees of Freedom of **Residual/Error Term** in **full model**

df_r : Degees of Freedom of **Residual/Error Term** in **restricted model**

$$F = \frac{(R_f^2 - R_r^2)/(df_r - df_f)}{(1 - R_f^2)/df_f} \sim F_{(df_r - df_f), df_f}$$

Full Model	
R Square	82.20%
Degrees of Freedom	24
Small Model	
R Square	76.77%
Degrees of Freedom	26

	Value	Df
Numerator	0.0272	2
Denominator	0.0074	24
F-Statistic	3.663	
α	5%	
Critical Value	3.403	
Conclusion	Model Improvement	
p-value	4.09%	
Conclusion	Model Improvement	

- $R^2_f = 82.2\%, df_f = 24, R^2_r = 76.7\%, df_r = 26 \Rightarrow F$-statistic ≈ 3.663. $3.663 > F_{2,24,0.95} \approx 3.403 \Rightarrow F$-statistic observation in 5% tail of $F \sim F_{2,24}$

  Conclusion: Reject H_0 in favor of H_1 **and an improvement in the model is detected in terms of the increased R^2-value**.

- This test requires acceptance of **normality assumption of residuals** for both models!

- This test is usefull when small increases in R^2 are observed, but **it requires** that **the smaller model to be nested in the larger model**.

However, **the improvement in R^2 and reduction in the Standard Error (SE) here comes at a cost**! Note the increase in VIF Factors of the coefficients:

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.146820	82.20%	77.76%	69.05%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	1.424	0.127	11.19	0.000	
Elevation	0.0483	0.0286	1.69	0.104	21.57
Sewer	-0.000048	0.000012	-3.86	0.001	1.32
Date	0.00548	0.00135	4.06	0.000	1.52
Flood	-0.394	0.102	-3.88	0.001	2.00
County	-0.113	0.110	-1.03	0.312	4.11
County*Elevation	-0.0307	0.0297	-1.03	0.312	28.12

Regression Equation

Log(Price) = 1.424 + 0.0483 Elevation - 0.000048 Sewer + 0.00548 Date - 0.394 Flood - 0.113 County - 0.0307 County*Elevation

Durbin-Watson Statistic

Durbin-Watson Statistic = 2.11482

Conclusion: While the standard error in the residual error term has reduced from 0.1611 to 0.1468, the uncertainty in standard errors of the coefficient estimators have increased.

- **Both the standard errrors** in the coefficients and the residuals contribute to **the standard error of the prediction, i.e. its uncertainty.** For the 247 acres property at hand we have:

Prediction for Log(Price)

Regression Equation

Log(Price) = 1.424 + 0.0483 Elevation - 0.000048 Sewer + 0.00548 Date - 0.394 Flood - 0.113 County - 0.0307 County*Elevation

Settings

Variable	Setting
Elevation	0
Sewer	0
Date	0
Flood	0
County	1
County*Elevation	0

Prediction

Fit	SE Fit	95% CI	95% PI
1.31073	0.106580	(1.09076, 1.53070)	(0.936284, 1.68518)

Prediction interval width is now: $1.68518 - 0.936284 \approx 0.748896$

- Prediction of Log(Price) using **the smaller model:**

Prediction for Log(Price)

Regression Equation

Log(Price) = 1.4891 + 0.01411 Elevation - 0.000044 Sewer + 0.00741 Date - 0.3183 Flood

Settings

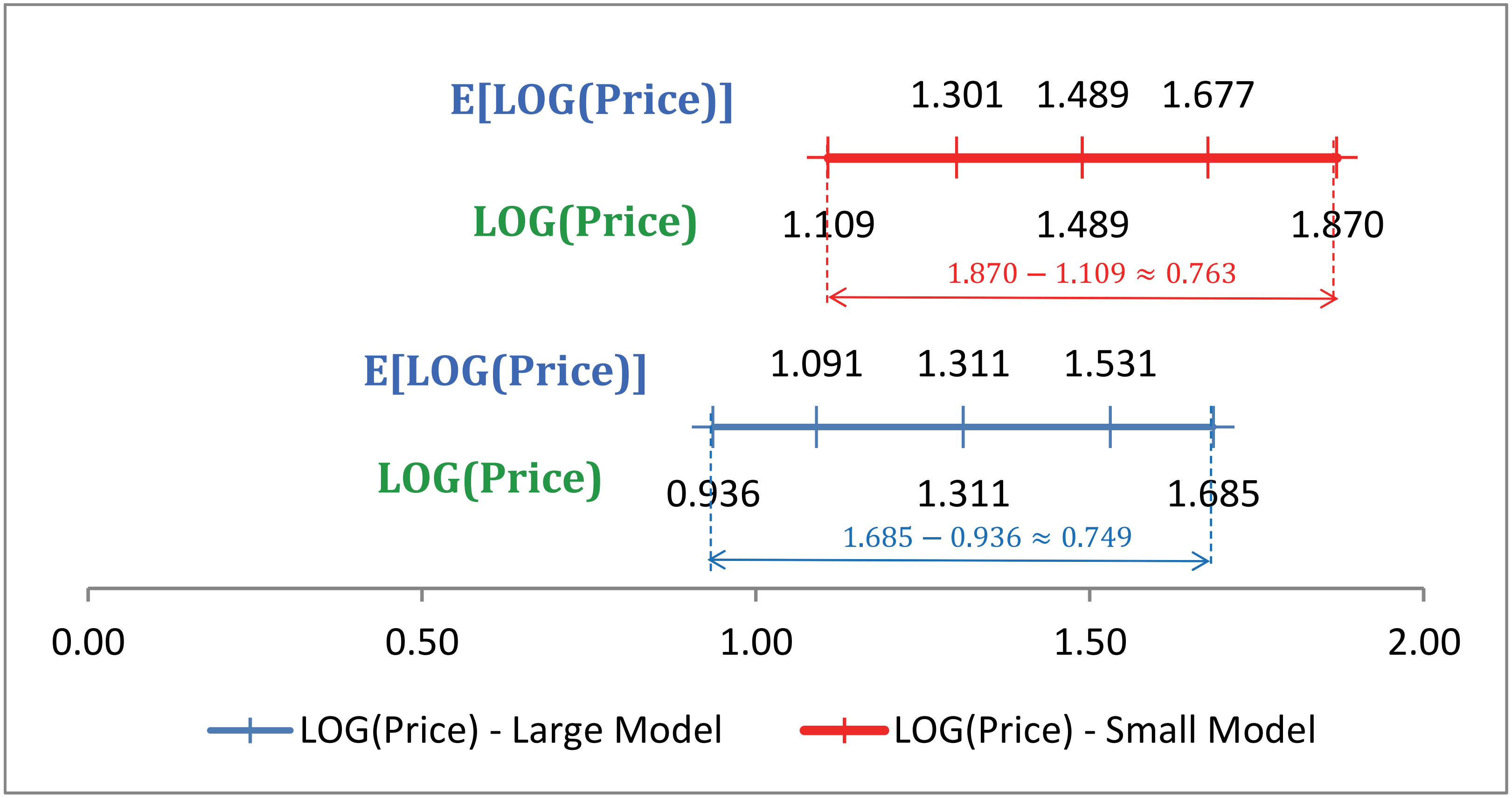
Variable	Setting
Elevation	0
Sewer	0
Date	0
Flood	0

Prediction

Fit	SE Fit	95% CI	95% PI
1.48907	0.0914849	(1.30102, 1.67712)	(1.10816, 1.86999)

Prediction interval width was: $1.86999 - 1.10816 \approx 0.76183$

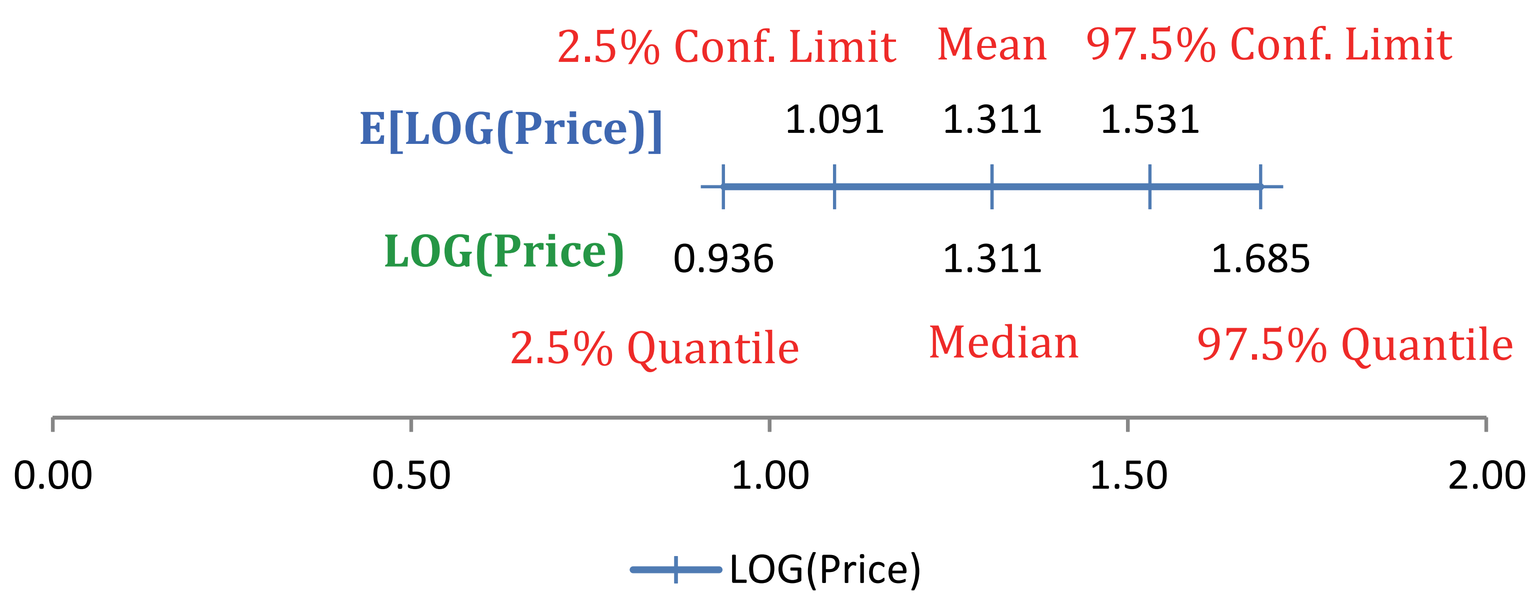
Conclusion: Prediction interval width of the smaller model is larger than the prediction interval width of the full model **despite the large VIF factors.** **Thus continue to predict\forecast with the full\larger model!**



Conclusion: Prediction interval width of the smaller model is larger than the prediction interval width of the full model **despite the large VIF factors.**

Thus continue to predict\forecast with the full\larger model!

Fit	SE Fit	95% CI	95% PI
1.31073	0.106580	(1.09076, 1.53070)	(0.936284, 1.68518)



Observe the Median and the Mean are of the same value! Why?

$$Y = \boldsymbol{x}_0^T \widehat{\boldsymbol{b}} + \epsilon,\ E[\epsilon] = 0, \epsilon \sim \boldsymbol{N(0,\sigma)} \Leftrightarrow E[Y|\boldsymbol{x}_0] = \boldsymbol{x}_0^T \widehat{\boldsymbol{b}}$$

Prediction

Fit	SE Fit	95% CI	95% PI
1.31073	0.106580	(1.09076, 1.53070)	(0.936284, 1.68518)

$$Pr(Log(Price) \leq 1.31073 | x_0) \approx 50\% \Leftrightarrow Pr(10^{Log(Price)} \leq 10^{1.31073} | x_0) \approx 50\%$$

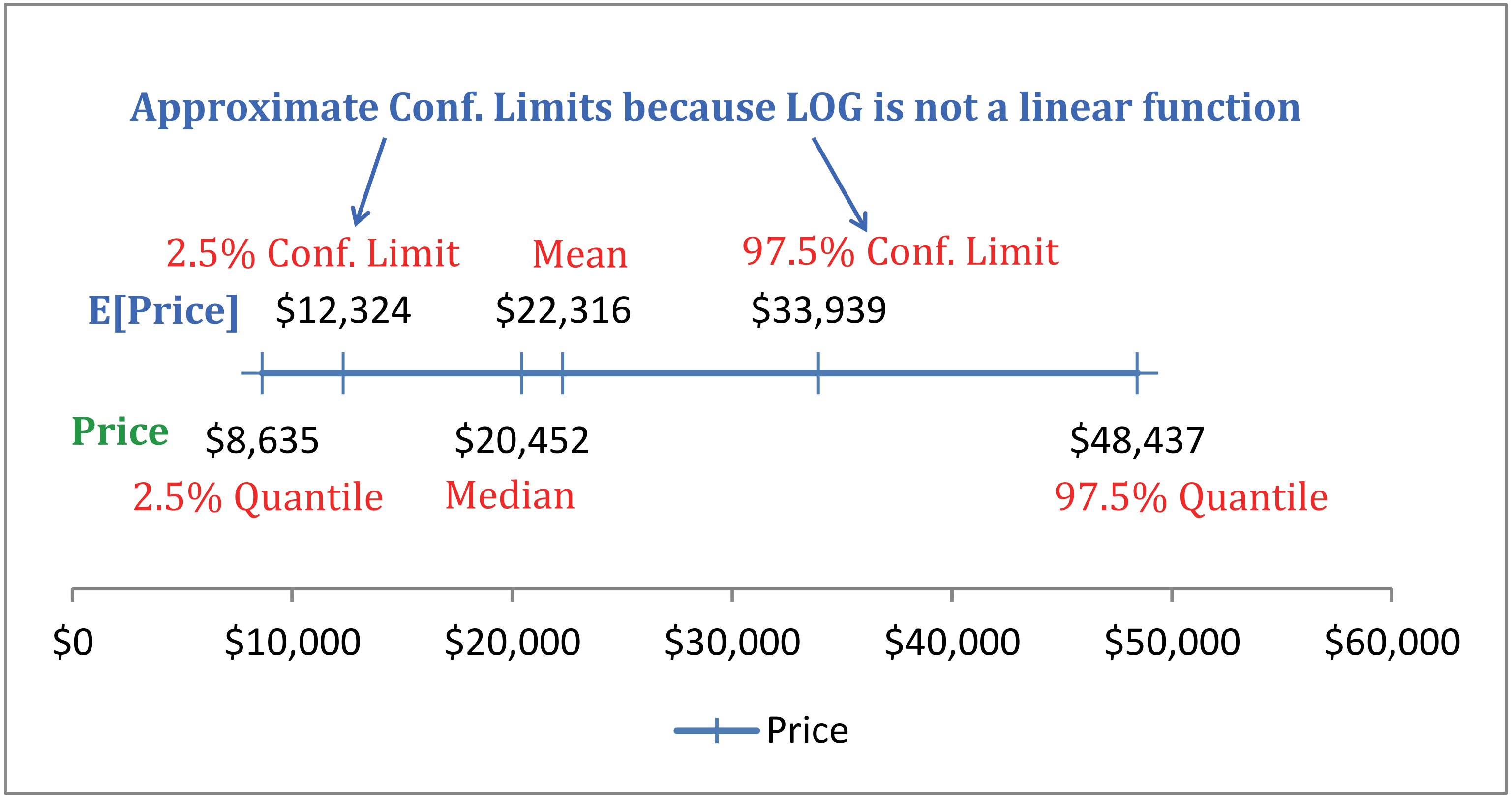
$$\Leftrightarrow Pr(Price \leq \$20452 | x_0) \approx 50\% \text{ (Recall } Price \text{ was measured in \$000's)}$$

Hence: $20452 is **the median estimate for the Price per Acre**

Thus we have here that: $Med[Log(Price)] = Log(Med[Price])$

95% Confidence Interval	
LB E[LOG(PRICE)]	1.09076
UB E[LOG(PRICE)]	1.53070
Approximate 95% Confidence Interval	
LB E[PRICE]	$12,324.22
UB E[PRICE]	$33,939.08

95% Prediction Interval (or Credibility Interval)			
LB LOG(PRICE)	0.936284		
UB LOG(PRICE)	1.685175		
95% Prediction Interval (or Credibility Interval)			
PRICE	$8,635.44		
PRICE	$48,436.78		



These are approximate Confidence Limits since we know:

$$E[Log(Price)] \neq Log(E[Price])$$

How do we get? $\widehat{E}[Price|\boldsymbol{x_0}] \approx \$22,316$

Prediction

Fit	SE Fit	95% CI	95% PI
1.31073	0.106580	(1.09076, 1.53070)	(0.936284, 1.68518)

$$Y = \boldsymbol{x}_0^T \widehat{\boldsymbol{b}} + \epsilon,\ E[\epsilon] = 0, \epsilon \sim N(0, SE) \Rightarrow E[Y|\boldsymbol{x}_0] = \boldsymbol{x}_0^T \widehat{\boldsymbol{b}}$$

$$V[Y|\boldsymbol{x}_0] = V[\boldsymbol{x}_0^T \boldsymbol{b}] + V[\epsilon] \Rightarrow V[Y] = (SE_{Fit})^2 + (SE_{Residuals})^2$$

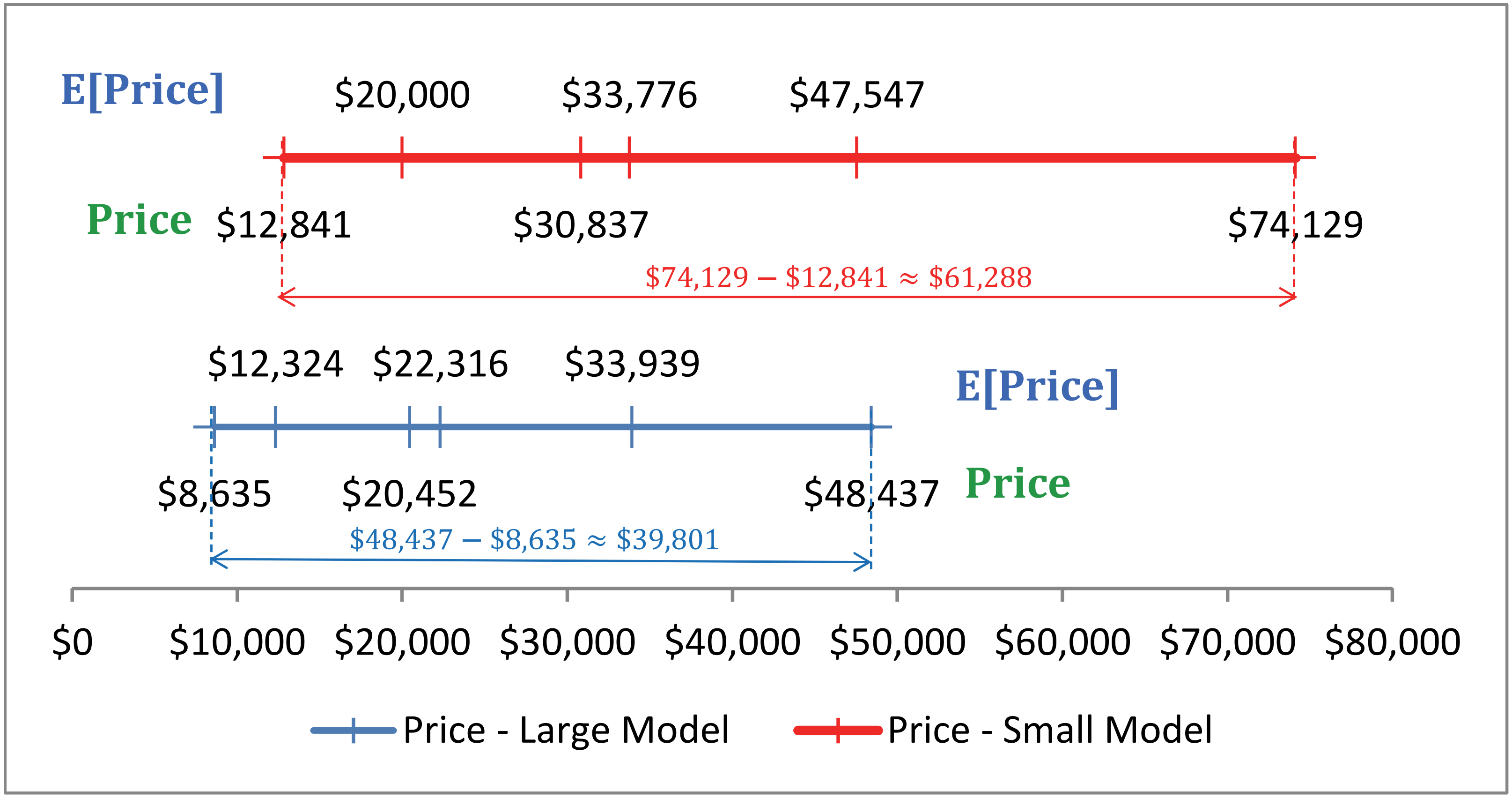
$$= (\mathbf{0.106580})^2 + (\mathbf{0.14682})^2 = (\mathbf{0.181426})^2$$

Summarizing:

$$\begin{cases} E[Log(Price)|\boldsymbol{x}_0] = 1.31073 = \mu, \\ V[Log(Price)|\boldsymbol{x}_0] = (\mathbf{0.181426})^2 = \sigma^2 \end{cases} \Rightarrow E[Price] = 10^{\mu + Ln(10) \times \frac{\sigma^2}{2}}$$

$$E[Price] \approx 10^{1.311 + 2.302 \times \frac{(0.181)^2}{2}} \approx 22.316\ (\times \$1000 \approx \$22,316).$$

Note the following formula in the book is wrong: $E[Price] = 10^{\mu + \frac{\sigma^2}{2}}$



Conclusion: Prediction interval width of the smaller model is larger than the prediction interval width of the full model **despite the large VIF factors.**

Thus predict\forecast with the full\larger model!

Prediction

Fit	SE Fit	95% CI	95% PI
1.31073	0.106580	(1.09076, 1.53070)	(0.936284, 1.68518)

- The variance of the term $\boldsymbol{x}_0^T \boldsymbol{b}$, where $\boldsymbol{b}$ is the estimator vector of the coefficients, is estimated by $\sqrt{SE^2 \times \boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0}$ and is referred to as **the sampling error.** **SE is the standard error of the residuals.**

- **A 100$(1-\alpha)$% confidence interval for the mean $E[y|\boldsymbol{x}_0]$ is :**

$$\boldsymbol{x}_0^T \widehat{\boldsymbol{b}} \pm t_{n-p-1,1-\alpha/2} \times SE \times \sqrt{[\boldsymbol{x}_0^T (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_0]}$$

$$\boldsymbol{x}_0^T \widehat{\boldsymbol{b}} \pm t_{n-p-1,1-\alpha/2} \times (0.106580) = (1.09076, 1.53070)$$

Standard Error : sample standard deviation of the residuals.
$\boldsymbol{x}_0$: values of the explanatory variables for which you would to forcast the dependent variable y.
$\widehat{b}$: The estimates of the regression coefficients.

Prediction

Fit	SE Fit	95% CI	95% PI
1.31073	0.106580	(1.09076, 1.53070)	(0.936284, 1.68518)

- The variance of the residuals SE^2 is referred to as **the model error**. The variance of the prediction $Y = \boldsymbol{x}_0^T\boldsymbol{b} + \epsilon$ is **the sum of the sampling error and the model error (there is independence between the two terms).**

$$Var(Y|x_0) = \textbf{(Standard Error)}^2 \times [\boldsymbol{x}_0^T\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0\,] + \textbf{(Standard Error)}^2$$
$$= (\mathbf{0.106580})^2 + (\mathbf{0.14682})^2 = (\mathbf{0.181426})^2$$

- **A 100$(1-\alpha)$% prediction interval for the random variable $(y|x_0)$ is :**

$$\boldsymbol{x}_0^T\,\widehat{\boldsymbol{b}} \pm t_{n-p-1,1-\alpha/2} \times \sqrt{Var(Y|x_0)} = (0.936284, 1.68518)$$

$$\boldsymbol{x}_0^T\,\widehat{\boldsymbol{b}} \pm t_{n-p-1,1-\alpha/2} \times (0.181426) = (0.936284, 1.68518)$$

$\boldsymbol{x}_0$: values of the explanatory variables to forcast the dependent variable y.

$\widehat{\boldsymbol{b}}$: The estimates of the regression coefficient.

- **Summarizing,** the value $\widehat{y} = \boldsymbol{x}_0^T \widehat{\boldsymbol{b}}$ is both an estimate of **the random variable $(Y|\boldsymbol{x}_0)$**, but also of **its expected value $E[Y|\boldsymbol{x}_0]$.**

- When describing the uncertainty in the estimate for $E[Y|\boldsymbol{x}_0]$, one only has to account for **the uncertainty in the regression coefficients**, leading to **a confidence interval** for $E[Y|\boldsymbol{x}_0]$.

- When describing the uncertainty in the random variable $(Y|x_0)$ one has to account for both **the uncertainty in the regression coefficients** and **in the residuals**, leading to **a prediction/credibility interval** for $(y|x_0)$.

- The vector $\boldsymbol{b}$ is an **estimator -vector** for the regression coefficients, where $\boldsymbol{b} \sim \boldsymbol{MVN}(\widehat{\boldsymbol{b}}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. For its variance-covariance matrix estimate we have $\widehat{\Sigma}(\boldsymbol{b}) = SE^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, where SE is **the residual standard error**.

- From the above it follows that:

$$V\widehat{ar}(Y|\boldsymbol{x}_0]) = SE^2 \times \left[\boldsymbol{x}_0^T\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_0 + 1\right] = (\mathbf{0.181426})^2$$

- It is known **for the natural logarithm $Ln(\cdot)$** that:

$$X \sim LN(\mu, \sigma) \Leftrightarrow Ln(X) \sim N(\mu_x, \sigma_x) \Rightarrow E[X|\mu_x, \sigma_x] = e^{\mu_x + \sigma_x^2/2} \quad (1)$$

- It is known for **the logarithm $Log(\cdot)$ with base 10** that:

$$Log(Y) = \frac{Ln(Y)}{Ln(10)} \Leftrightarrow Ln(Y) = Ln(10) \times Log(Y) \quad (2)$$

- Hence from (2), **$Ln(Y)$ is linear transformation of $Log(Y)$** :

$$Log(Y) \sim N(\mu, \sigma) \Rightarrow Ln(Y) \sim N\{Ln(10) \times \mu, Ln(10) \times \sigma\} \quad (3)$$

- With **the expected value expression** in (1) and (3) it now follows that:

$$\begin{cases} \mu_y = Ln(10) \times \mu \\ \sigma_y = Ln(10) \times \sigma \end{cases} \Rightarrow \begin{cases} E[Y|\mu, \sigma] = e^{\mu_y + \sigma_y^2/2} = e^{Ln(10) \times \mu + \{Ln(10) \times \sigma\}^2/2} \\ \qquad = \left[e^{Ln(10)}\right]^{\mu + Ln(10) \times \sigma^2/2} \\ \qquad = 10^{\mu + Ln(10) \times \sigma^2/2} \end{cases} \quad (4)$$