# STAT021 Statistical Methods II

## Lecture 11 SLR Prediction

Lu Chen
Swarthmore College
10/9/2018

# Review - Simple Linear Regression

**CHOOSE**

- Exploratory data analysis; Model: $Y = \beta_0 + \beta_1 X + \epsilon$ where $\epsilon \overset{iid}{\sim} N(0, \sigma)$

**FIT**

- Maximum likelihood estimation (MLE)

**ASSESS model**

- Inference for the intercept and slope; ANOVA and $R^2$

**ASSESS error**

- Check conditions and transformations; Outliers and influential points

**USE**

- Predictions

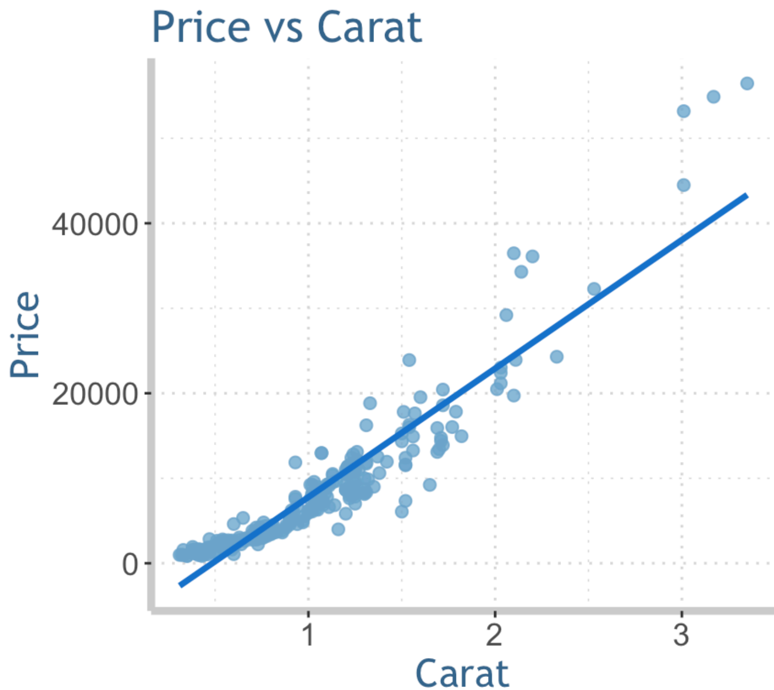
# Review - Simple Linear Regression

- Simple linear regression ANOVA
  - Sum of squares and degree of freddom
  - Mean square, $F$ test and $R^2$
  - ANOVA table
- Regression and correlation
  - $t$ test for correlation
- Three tests for linear relationship?
- Transformation
  - Example 1: Diamond price
  - Example 2: Valentine's Day love level

# Outline

- Example 1: Diamond price
- Prediction
  - Mean response
  - Individual response
- Inference for predictions
  - Confidence interval for a mean response
  - Prediction interval for an individual response
- Prediction after transformation
  - Transformation and transforming back
- Example 2: Valentine's Day love level
- Example 3: UK dog food volume by year

# Example 1: Diamond price




**Estimated regression line:**

$$\hat{y} = -7342 + 15130x$$

$$\widehat{Price} = -7342 + 15130 \times Carat$$

- For $Carat = 1.5$, what's the value of $Price$?
- $\widehat{Price} = -7342 + 15130 \times 1.5 = 15353$
- What does this value mean?
- 1. For diamonds of 1.5 carats, their average price is predicted as $15,353.
- 2. For a 1.5-carat diamond, its price is predicted as $15,353.

# Prediction - Mean response

$$
\begin{array}{rccccccl}
 & \text{Data} & = & \text{Model} & + & \text{Error} & & \\
\text{Population:} & Y & = & \mu_Y & + & \epsilon & & \text{where } \epsilon \overset{iid}{\sim} N(0, \sigma) \\
 & Y & = & \beta_0 + \beta_1 X & + & \epsilon & & \\
\text{Sample:} & y & = & b_0 + b_1 x & + & e & &
\end{array}
$$

- $\mu_Y = \beta_0 + \beta_1 X$
- For a given $x^*$ value that we are interested in, the predicted **mean response** is

$$\boldsymbol{\hat{\mu}_y = b_0 + b_1 x^*}$$

- For the diamond example, average price of diamonds of $x^* = 1.5$ carats is $\hat{\mu}_{Price} = -7342 + 15130 \times 1.5 = 15353$

# Prediction - Individual response

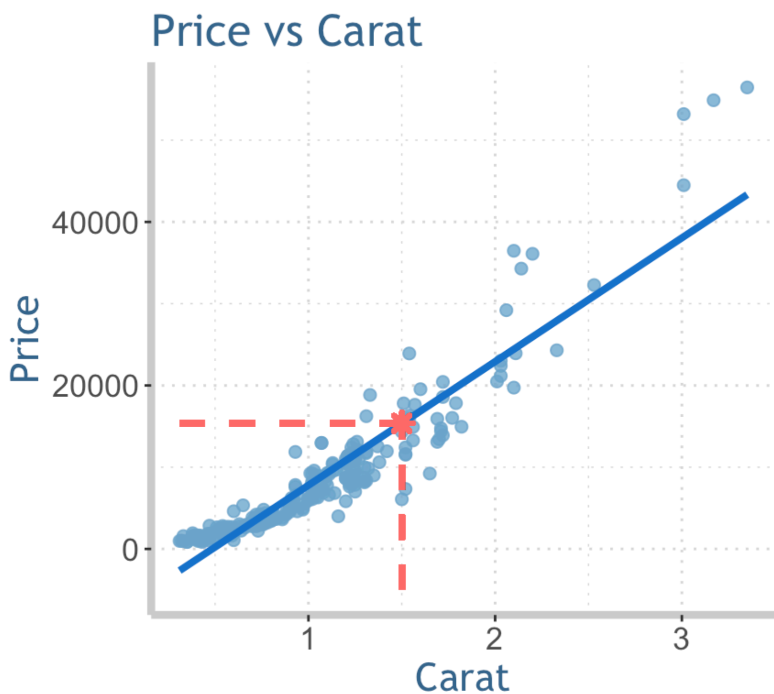
$$
\begin{array}{rcccccl}
 & \text{Data} & = & \text{Model} & + & \text{Error} & \\
\text{Population:} & Y & = & \mu_Y & + & \epsilon & \text{where } \epsilon \overset{iid}{\sim} N(0, \sigma) \\
 & Y & = & \beta_0 + \beta_1 X & + & \epsilon & \\
\text{Sample:} & y & = & b_0 + b_1 x & + & e &
\end{array}
$$

- $e$ is a random number following Normal distribution, for a specific prediction given $x^*$ value, we do not know the value of $e$.
- Therefore, the best prediction we can have for an **individual response** is

$$\boldsymbol{\hat{y} = b_0 + b_1 x^*}$$

- For the diamond example, the predicted price for a diamond of $x^* = 1.5$ carats is $\widehat{Price} = -7342 + 15130 \times 1.5 = 15353$

# Mean response & individual response




- The **mean response** $\mu_y = \beta_0 + \beta_1 x^*$ is predicted as $\hat{\mu}_y = b_0 + b_1 x^*$.
- The **individual response** $y = \beta_0 + \beta_1 x^* + \epsilon$ is predicted as $\hat{y} = b_0 + b_1 x^*$.
- They have the same value.
- **What's the difference?**
- When predicting the mean response, the uncertainty comes from $b_0$ and $b_1$, which are estimated from sample data.
- When predicting the individual response, the uncertainty comes from $b_0$, $b_1$, and the error term.

# Variability of a mean and an invidual response

The variability of an estimated mean response $\hat{\mu}_y = b_0 + b_1 x^*$ is measured by

$$SE_{\hat{\mu}_y} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

For an estimated individual response $\hat{y} = b_0 + b_1 x^*$

$$SE_{\hat{y}} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x - \bar{x})^2}}$$

- $SE^2_{\hat{y}} = SE^2_{\hat{\mu}_y} + \hat{\sigma}^2 \Rightarrow SE_{\hat{y}} > SE_{\hat{\mu}_y}$.
- When the $x^*$ value is far away from the center $\bar{x}$, predictions will have large variability. Therefore, we should avoid **extrapolation**, i.e. predictions with $x^*$ values outside the range of the observed $x$ values.

# Confidence interval & prediction interval

A level $C$ **confidence interval** for a mean response $\mu_y$ and a level $C$ **prediction interval** for an individual response $y$ when $x$ takes value $x^*$ are

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}_y}, \qquad \hat{y} \pm t^* SE_{\hat{y}}$$

where

$$SE_{\hat{\mu}_y} = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x-\bar{x})^2}}, \quad SE_{\hat{y}} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x-\bar{x})^2}}$$

and $t^*$ is the value for the $t(n-2)$ density curve with area $C$ between $-t^*$ and $t^*$.

- The variability of a mean response is always smaller than that of an individual response for the same given $x^*$.
- A level $C$ confidence interval is always narrower than a level $C$ prediction interval for the same given $x^*$.

# Confidence interval & prediction interval

```
diaSLR <- lm(Price ~ Carat, data=Diamonds)
predict(diaSLR, list(Carat=1.5))
```

```
##       1
## 15353.5
```

```
predict(diaSLR, list(Carat=1.5), interval="confidence")
```

```
##       fit     lwr     upr
## 1 15353.5 14883.5 15823.5
```

- $\hat{\mu}_y = 15354$ with 95% CI $[14884, 15824]$

```
predict(diaSLR, list(Carat=1.5), interval="prediction")
```

```
##       fit      lwr      upr
## 1 15353.5 9706.664 21000.34
```

- $\hat{y} = 15354$ with 95% PI $[9707, 21000]$

```
predict(diaSLR, list(Carat=1.5), interval="prediction", level=0.99)
```

```
##       fit      lwr      upr
## 1 15353.5 7915.215 22791.79
```

- $\hat{y} = 15354$ with 99% PI $[7915, 22792]$

# Confidence interval & prediction interval

```r
# Predicted mean and individual response given all the x values in the data
ci <- predict(diaSLR, interval="confidence"); head(ci)
```

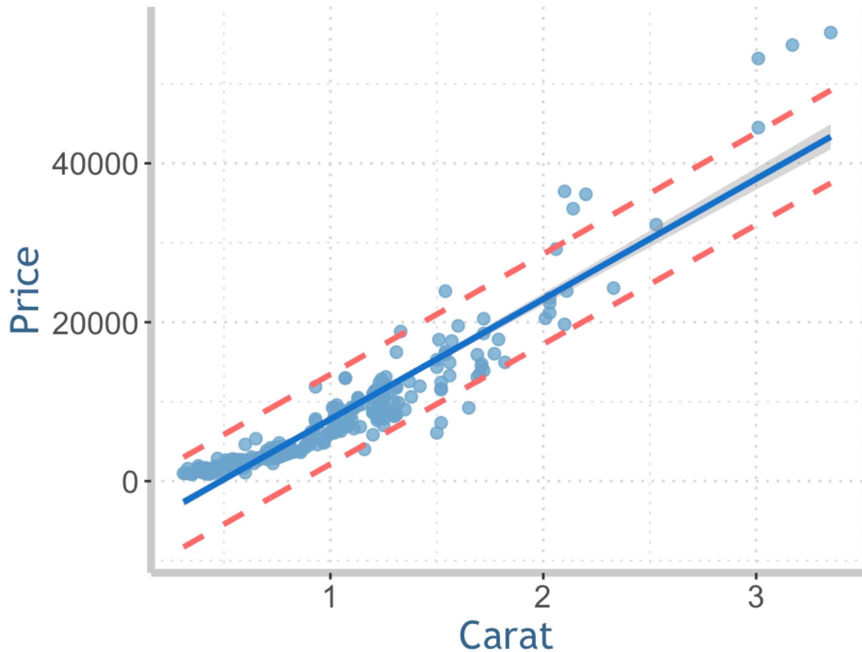
```
##         fit       lwr       upr
## 1  8998.842  8670.218  9327.466
## 2 -2651.368 -3189.549 -2113.187
## 3 -2500.066 -3033.035 -1967.097
## 4 -2348.765 -2876.551 -1820.978
## 5 -2348.765 -2876.551 -1820.978
## 6 -2046.162 -2563.673 -1528.651
```

```r
pi <- predict(diaSLR, interval="prediction"); head(pi)
```

```
##         fit       lwr       upr
## 1  8998.842  3362.010 14635.674
## 2 -2651.368 -8304.289  3001.554
## 3 -2500.066 -8152.494  3152.361
## 4 -2348.765 -8000.706  3303.176
## 5 -2348.765 -8000.706  3303.176
## 6 -2046.162 -7697.153  3604.829
```

# Confidence interval & prediction interval lines

```
Diamonds2 <- data.frame(Diamonds, pi)
ggplot(data=Diamonds2, aes(x=Carat, y=Price))+
  geom_point(color="skyblue3", size=2, alpha=0.8)+
  geom_smooth(method='lm', size=1.2, se=TRUE, color="dodgerblue3")+
  geom_line(aes(y=lwr), color="indianred1", linetype=2,size=1.1)+
  geom_line(aes(y=upr), color="indianred1", linetype=2,size=1.1)
```




- `se=TRUE` in `geom_smooth` adds the confidence interval lines.
- `geom_line` adds the prediction interval lines.
- Both confidence and prediction interval lines are not linear. The width of the intervals depends on value of $x$.

# Prediction after transformation

```r
diaSLR_new <- lm(log(Price) ~ log(Carat), data=Diamonds)
predict(diaSLR_new, list(Carat=1.5), interval="confidence") # log(Price)
```

```
##        fit      lwr      upr
## 1 9.488288 9.450453 9.526124
```

$\hat{\mu}_y = \hat{\mu}_{log(Price)} = 9.49$ with 95% CI $[9.45, 9.53]$

```r
predict(diaSLR_new, list(Carat=1.5), interval="prediction") # log(Price)
```

```
##        fit      lwr    upr
## 1 9.488288 9.053477 9.9231
```

$\hat{y} = \widehat{log(Price)} = 9.49$ with 95% PI $[9.05, 9.92]$

```r
exp(predict(diaSLR_new, list(Carat=1.5), interval="confidence")) # Price
```

```
##        fit      lwr      upr
## 1 13204.18 12713.92 13713.33
```

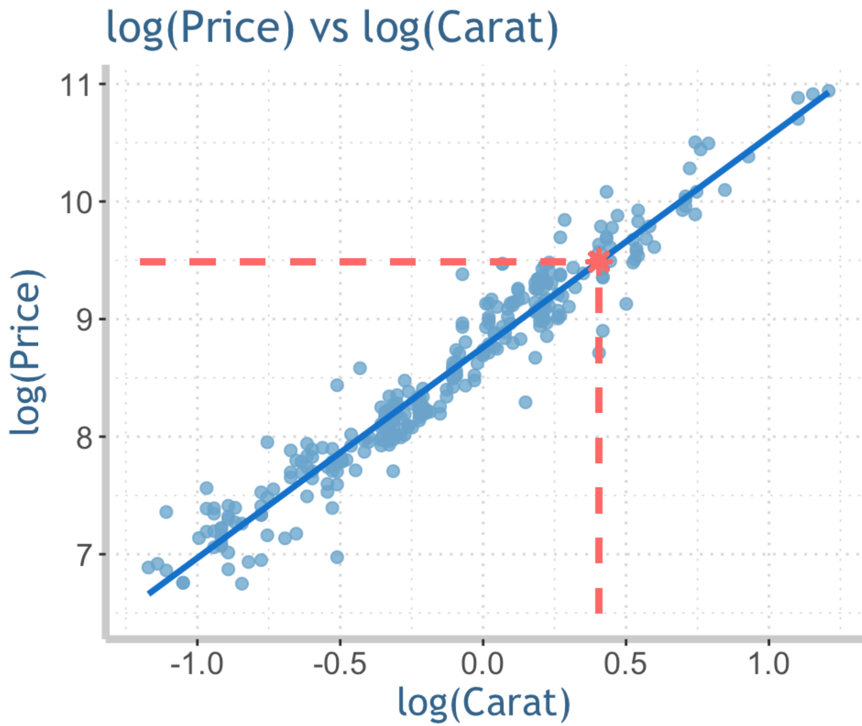
$\hat{\mu}_{Price} = e^{9.49} = 13204$ with 95% CI $[12714, 13713]$

```r
exp(predict(diaSLR_new, list(Carat=1.5), interval="prediction")) # Price
```

```
##        fit      lwr      upr
## 1 13204.18 8548.208 20396.12
```

$\widehat{Price} = e^{9.49} = 13204$ with 95% PI $[8548, 20396]$

# Prediction after transformation



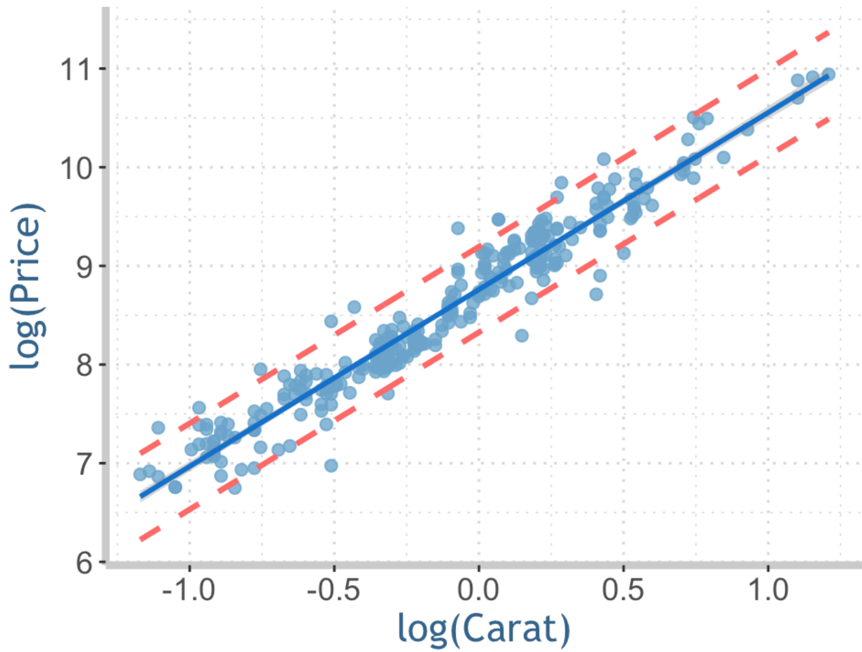

**Estimated regression line:**

$\widehat{\log(Price)} = 8.8 + 1.8 \times \log(Carat)$

- For $Carat = 1.5$, $\log(Price) = 9.49$
  - 95% confidence interval $[9.45, 9.53]$
  - 95% prediction interval $[9.05, 9.92]$
- $\widehat{Price} = e^{8.8+1.8\times\log(Carat)} = e^{9.49} = 13204$
  - The average price of 1.5-carat diamonds is predicted as \$13,204 with 95% confidence interval $[e^{9.45}, e^{9.53}] = [12714, 13713]$
  - The price of a 1.5-carat diamond is predicted as \$13,204 with 95% prediction interval $[e^{9.05}, e^{9.92}] = [8548, 20396]$

# Prediction after transformation

```
Diamonds3 <- data.frame(Diamonds, predict(diaSLR_new, interval="prediction"))
ggplot(data=Diamonds3, aes(x=log(Carat), y=log(Price)))+
  geom_point(color="skyblue3", size=2, alpha=0.8)+
  geom_smooth(method='lm', size=1, se=TRUE, color="dodgerblue3")+
  geom_line(aes(y=lwr), color="indianred1", linetype=2,size=1.1)+
  geom_line(aes(y=upr), color="indianred1", linetype=2,size=1.1)
```




$$\widehat{\log(Price)} = 8.8 + 1.8 \times \log(Carat)$$

- *Price* and *Carat* are displayed in the transformed scale.
- Let's transform it back to the original scale to display the relationship between *Price* and *Carat* directly and compare this new model to the old model.

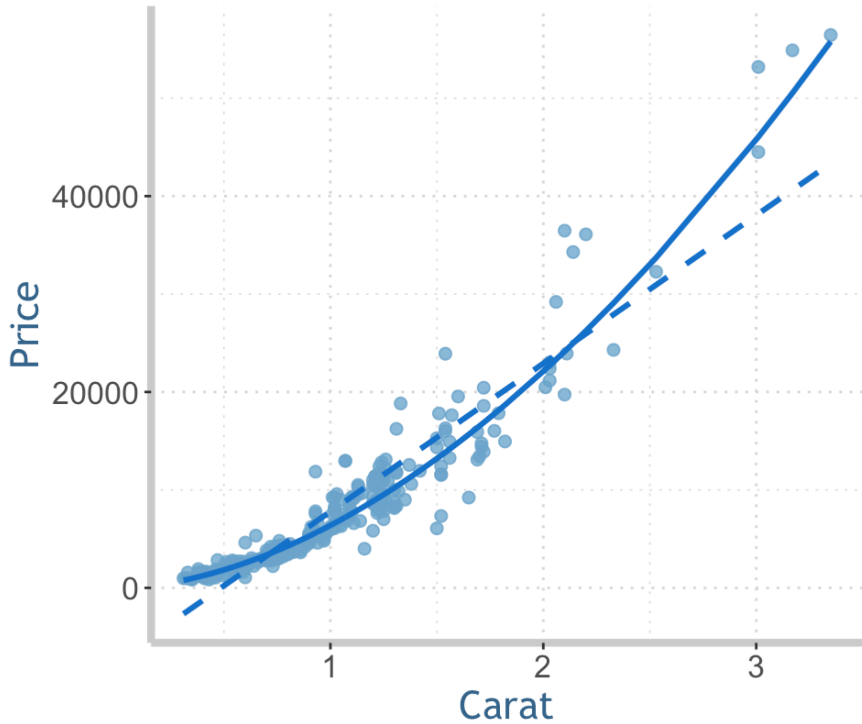
# Transformation and transforming back

```r
Diamonds4 <- data.frame(Diamonds,
    ci=predict(diaSLR, interval="confidence"), # CI based on old model
    pi=predict(diaSLR, interval="prediction"), # PI based on old model
    eci=exp(predict(diaSLR_new, interval="confidence")), # CI based on new model
    epi=exp(predict(diaSLR_new, interval="prediction"))) # PI based on new model
head(Diamonds4, 3)
```

```
##   Carat  Price    ci.fit    ci.lwr    ci.upr    pi.fit    pi.lwr
## 1  1.08 7228.8  8998.842  8670.218  9327.466  8998.842  3362.010
## 2  0.31  979.3 -2651.368 -3189.549 -2113.187 -2651.368 -8304.289
## 3  0.32 1010.9 -2500.066 -3033.035 -1967.097 -2500.066 -8152.494
##      pi.upr   eci.fit   eci.lwr   eci.upr   epi.fit   epi.lwr
## 1 14635.674 7325.9530 7129.2464 7528.0871 7325.9530 4746.4969
## 2  3001.554  781.2292  736.0640  829.1658  781.2292  504.5324
## 3  3152.361  826.9993  780.3679  876.4173  826.9993  534.2003
##     epi.upr
## 1 11307.199
## 2  1209.673
## 3  1280.284
```

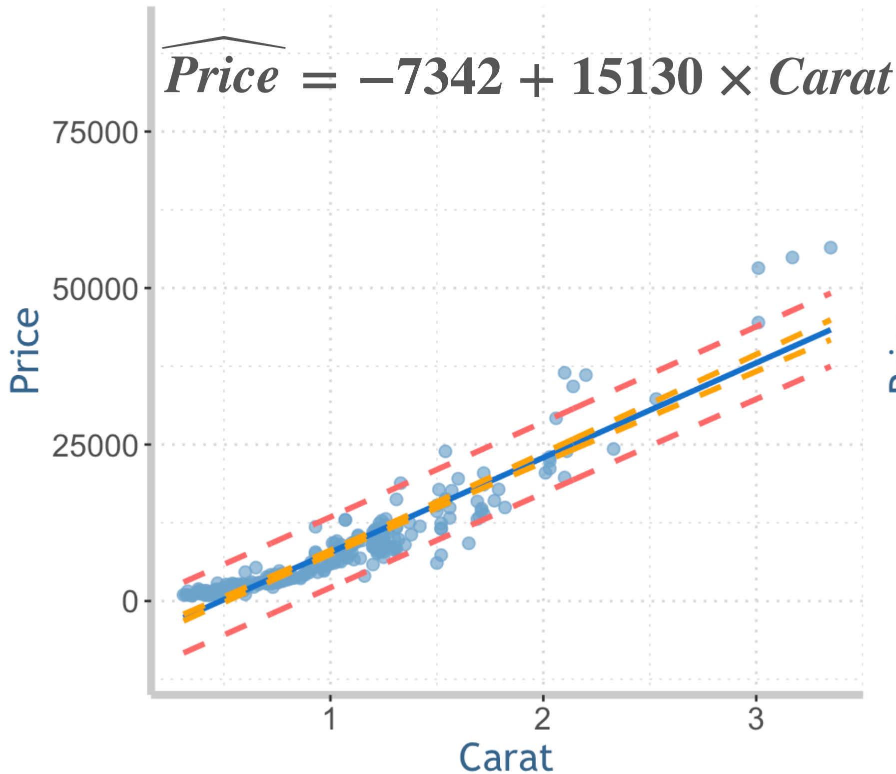
# Transformation and transforming back

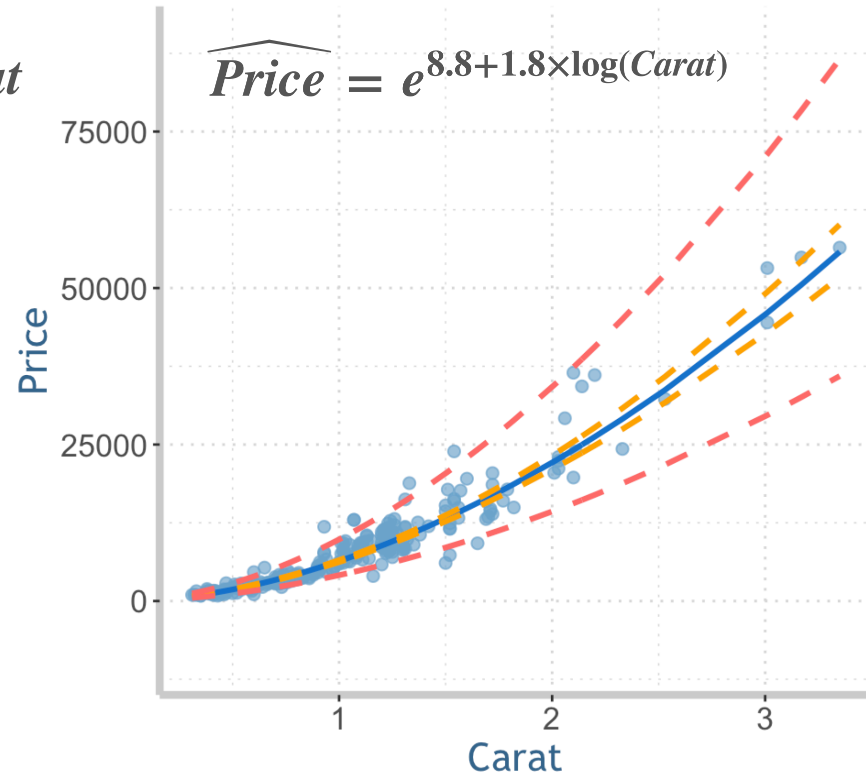
```
ggplot(data=Diamonds4, aes(x=Carat, y=Price))+
  geom_point(color="skyblue3", size=2, alpha=0.8)+
  geom_line(aes(y=ci.fit), color="dodgerblue3", size=1.1, linetype=2)+
  geom_line(aes(y=eci.fit), color="dodgerblue3", size=1.1)
```




- Compare the old model
  $\widehat{Price} = -7342 + 15130 \times Carat,$
  to the new model
  $\widehat{\log(Price)} = 8.8 + 1.8 \times \log(Carat)$ or
  $\widehat{Price} = e^{8.8+1.8\times\log(Carat)}$
  The new model captures the pattern of the data much better.
- To tranform the data is in fact to find a better way to describe the variability in the data.

# Transformation and transforming back




- Another benefit we get from the new model is that the values of *Carat*, *Price*, the confidence intervals and the prediction intervals are all bound to be greater than 0, which is more realistic.

# Transformation and transforming back

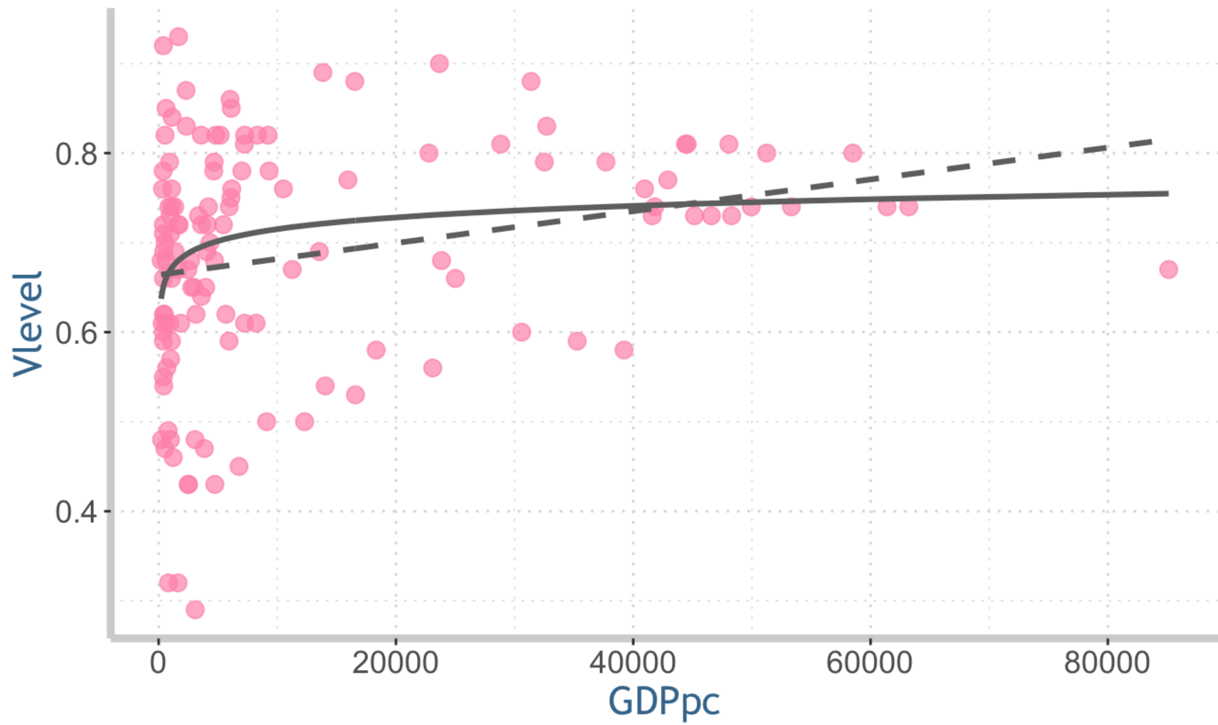
## R codes for adding 95% CI and PI lines after transformation

```
ggplot(data=Diamonds4, aes(x=Carat, y=Price))+
  geom_point(color="skyblue3", size=2, alpha=0.7)+
  geom_line(aes(y=ci.fit), color="dodgerblue3",size=1.1)+
  geom_line(aes(y=ci.lwr), color="orange", linetype=2,size=1.1)+
  geom_line(aes(y=ci.upr), color="orange", linetype=2,size=1.1)+
  geom_line(aes(y=pi.lwr), color="indianred1", linetype=2,size=1.1)+
  geom_line(aes(y=pi.upr), color="indianred1", linetype=2,size=1.1)+
  ggtitle("Model before transformation")

ggplot(data=Diamonds4, aes(x=Carat, y=Price))+
  geom_point(color="skyblue3", size=2, alpha=0.7)+
  geom_line(aes(y=eci.fit), color="dodgerblue3",size=1.1)+
  geom_line(aes(y=eci.lwr), color="orange", linetype=2,size=1.1)+
  geom_line(aes(y=eci.upr), color="orange", linetype=2,size=1.1)+
  geom_line(aes(y=epi.lwr), color="indianred1", linetype=2,size=1.1)+
  geom_line(aes(y=epi.upr), color="indianred1", linetype=2,size=1.1)+
  ggtitle("Model after transformation")
```

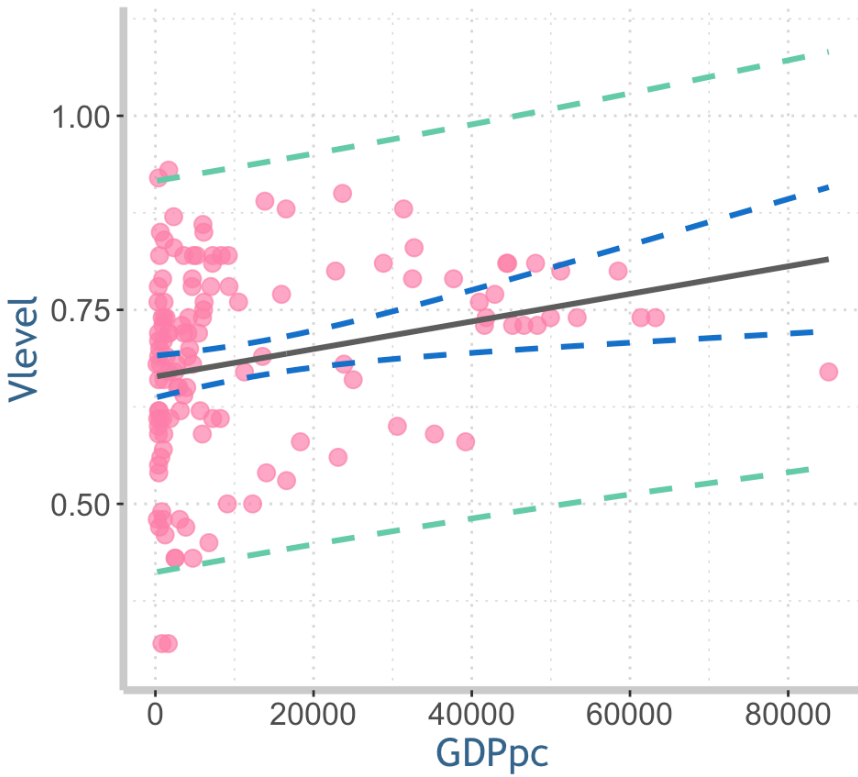
# Example 2: Valentine's Day love level

- **Old model** $\widehat{Vlevel} = 0.66 + 1.8 \times 10^{-6} \times GDPpc$
- **New model** $\widehat{Vlevel} = \sqrt{0.26 + 0.03 \times \log(GDPpc)}$ from $\widehat{Vlevel^2} = 0.26 + 0.03 \times \log(GDPpc)$

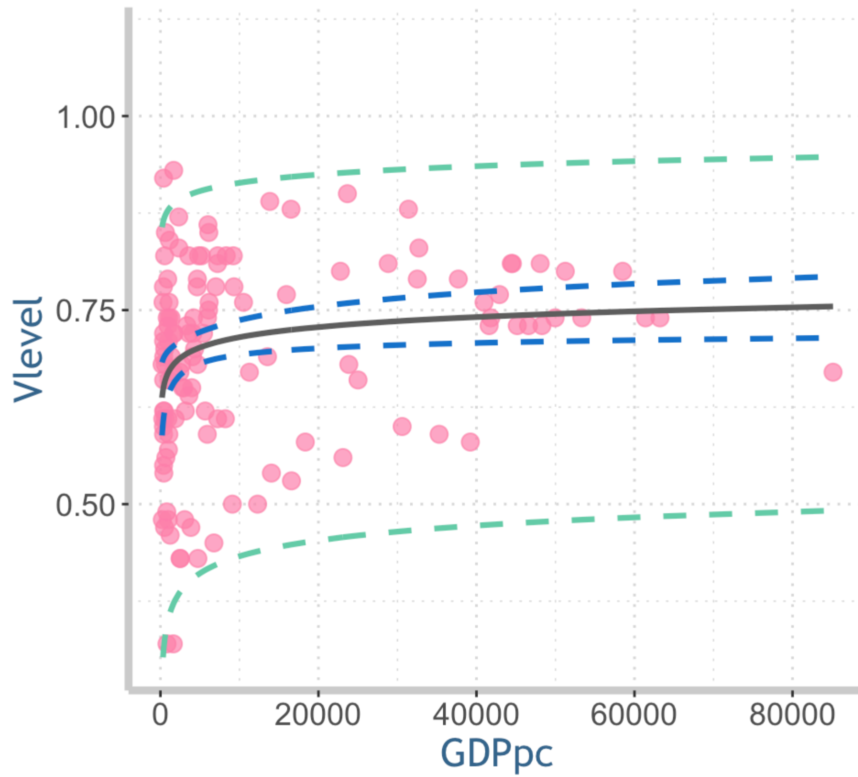


# Example 2: Valentine's Day love level

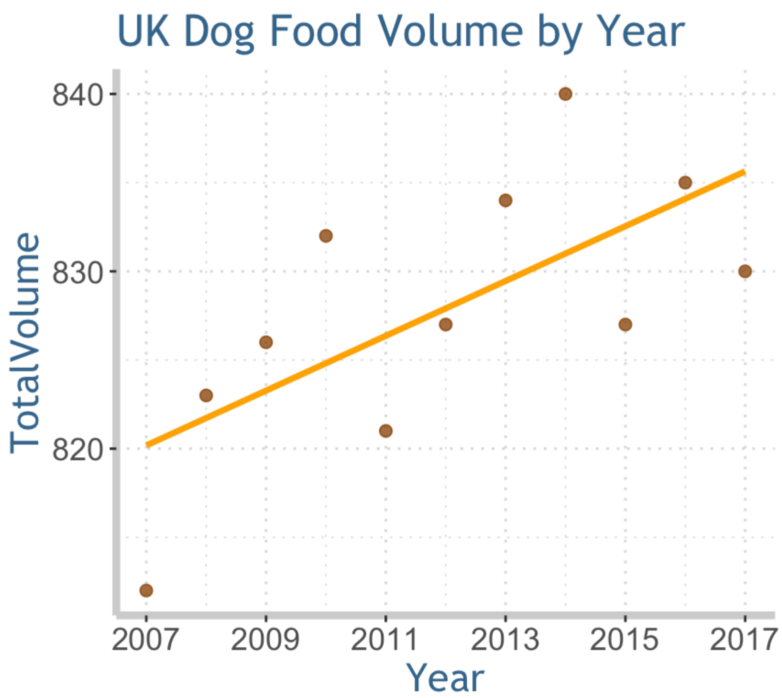
$$\widehat{Vlevel} = 0.66 + 1.8 \times 10^{-6} \times GDPpc$$




$$\widehat{Vlevel} = \sqrt{0.26 + 0.03 \times \log(GDPpc)}$$



# Example 3: UK dog food volume by year




**Estimated regression line:**

$$\widehat{TotalVolume} = -2281.5 + 1.5455 \times Year$$

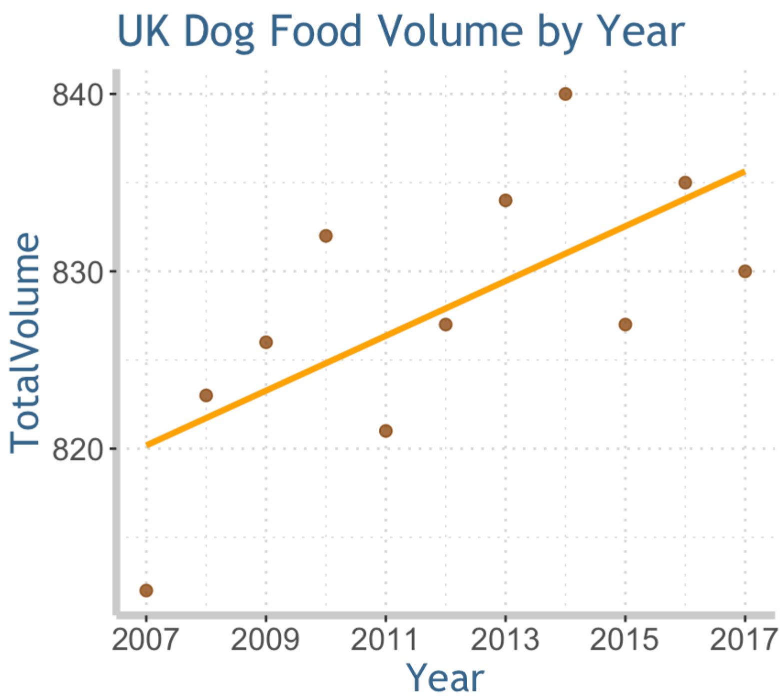
What's the predicted total volume for 2015?

- $\hat{y} = -2281.5 + 1.5455 \times 2015 = 832.7$

What's the residual of the prediction for 2015?

- The observed total volume for 2015 is 827.
- $e = y - \hat{y} = 827 - 832.7 = -5.7$

# Example 3: UK dog food volume by year




**Estimated regression line:**

$$\widehat{TotalVolume} = -2281.5 + 1.5455 \times Year$$

What's the predicted total volume and the corresponding 95% interval for 2018?

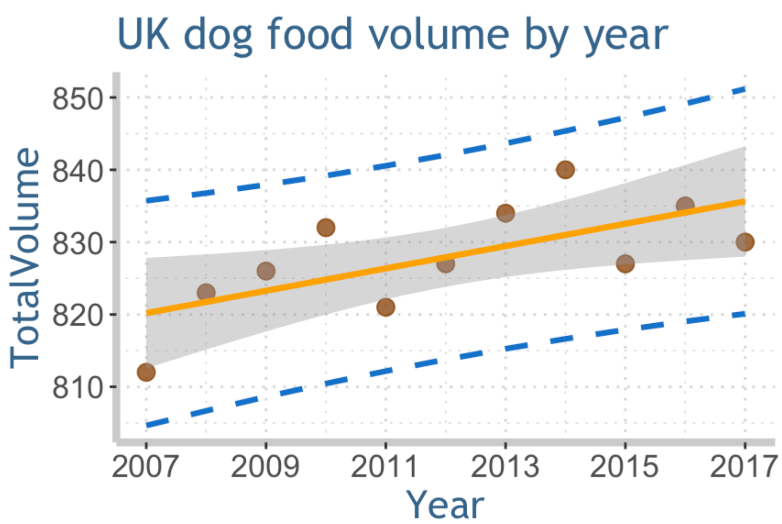
- $\hat{y} = -2281.5 + 1.5455 \times 2018 = 837.3$
- The 95% prediction interval is [821.1, 853.3].
- Here $837.3 \neq 837.2$ due to rounding errors.

```r
predict(dogmodel, list(Year=2018), interval="prediction")
```

```
##          fit      lwr      upr
## 1 837.1818 821.0636 853.3001
```

# Example 3: UK dog food volume by year

```r
dogfood2 <- data.frame(dogfood, predict(dogmodel, interval="prediction"))
ggplot(data=dogfood2, aes(x=Year, y=TotalVolume))+
  geom_point(color="darkorange4", size=3, alpha=0.8)+
  # Add regression line with 95% CI lines
  geom_smooth(method='lm', size=1.2, color="orange")+
  # Add 95% PI line
  geom_line(aes(y=lwr), color="dodgerblue3", linetype=2, size=1.1)+
  geom_line(aes(y=upr), color="dodgerblue3", linetype=2, size=1.1)+
  ggtitle("UK dog food volume by year")
```



# Summary

- Prediction
  - Mean response $\hat{\mu}_y$ and confidence interval
    - Smaller variability, narrower interval
  - Individual response $\hat{y}$ and prediction interval
    - Larger variability, wider interval
- Predictions using $x^*$ values far away from the center have larger variability than predictions using $x^*$ values close to the center.
- Transforming variables in a simple linear regression model allows us to model non-linear relationship and/or data with non-Normal error distributions.