STAT021 Statistical Methods II

Lecture 9 Simple Linear Regression

Lu Chen
Swarthmore College
10/2/2018

Review - ANOVA

- One-way ANOVA model and table
  - $Y = \mu + \alpha_k + \epsilon$, where $k = 1, 2, \cdots, K$ and $\epsilon \overset{iid}{\sim} N(0, \sigma)$
  - Sum of squares, degrees of freedom, mean square
- ASSESS model
  - F test and R^2
  - Multiple pairwise comparisons - control type I error rate
- ASSESS error
  - Zero mean, equal variance, Normality, independence
- Two way ANOVA model
  - Additive model $Y = \mu + \alpha_k + \beta_j + \epsilon$
  - Model with interaction $Y = \mu + \alpha_k + \beta_j + \gamma_{kj} + \epsilon$
  - Interaction plot

Outline - Simple Linear Regression

CHOOSE

- Exploratory data analysis; Model definition

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

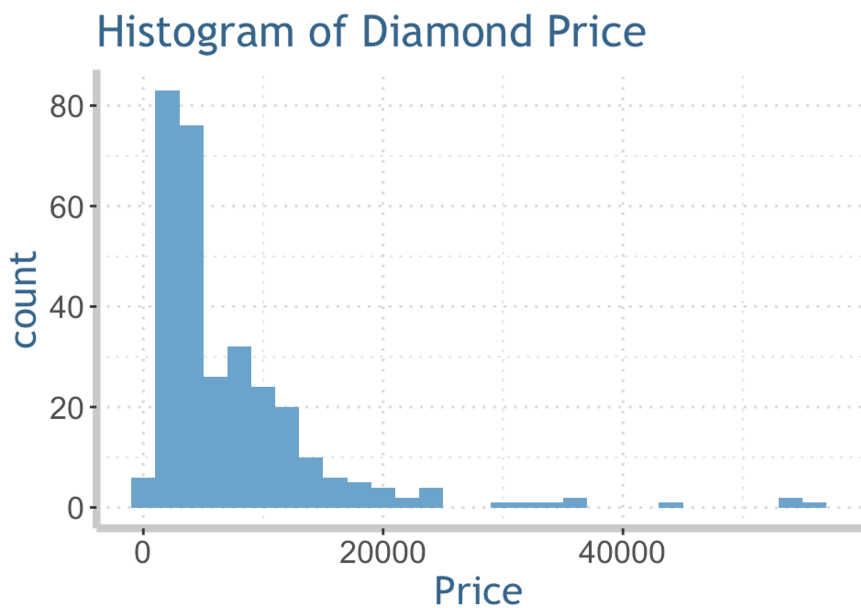
USE

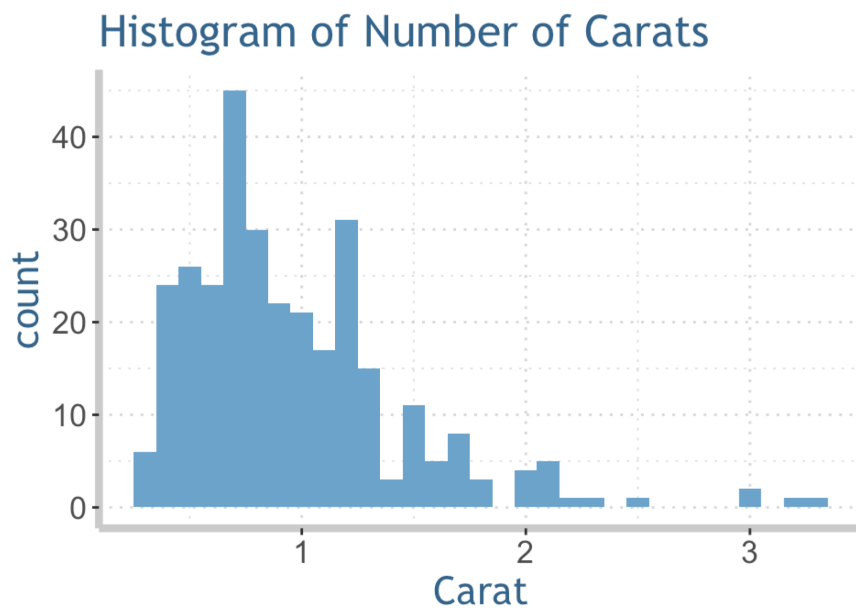
- Predictions

CHOOSE: exploratory data analysis

Diamond price and number of carats

- Response variable: *Price*, quantitative; mean \$7381.3, SD \$8000.3.
- Explanatory variable: *Carat*, quantitative; mean 0.97, SD 0.49.





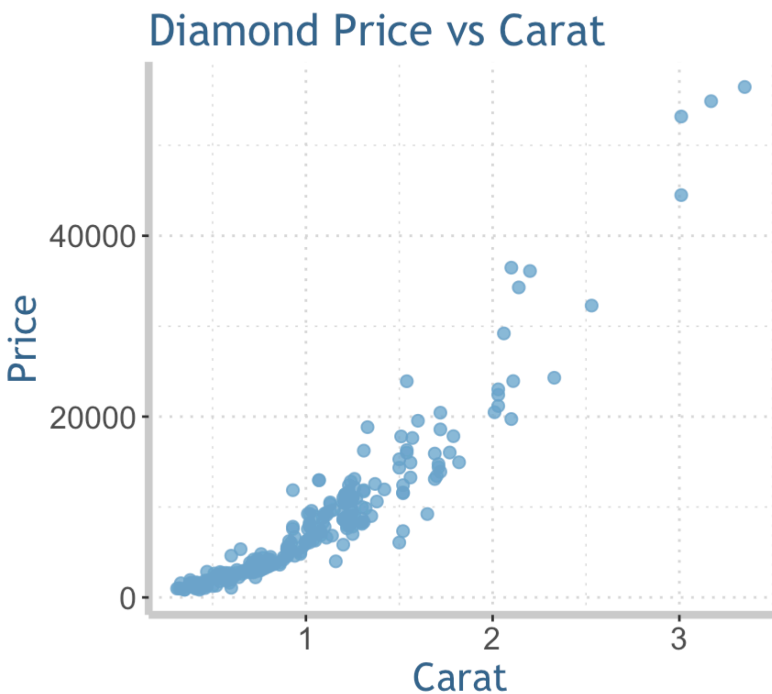
Both distributions are skewed to the right.

CHOOSE: exploratory data analysis

```
cor(Diamonds$Carat, Diamonds$Price) # correlation
```

```
## [1] 0.9341552
```

```
ggplot(data=Diamonds, aes(x=Carat, y=Price))+
  geom_point(color="skyblue3", size=2, alpha=0.8)+ # Scatterplot
  ggtitle("Diamond Price vs Carat")
```



- **Scatterplot** displays the relationship between two quantitative variables.
  - y-axis: response variable *Price*
  - x-axis: explanatory variable *Carat*
- Describe a scatterplot:
  - Form: linear or curved or none?
  - Direction: positive or negative?
  - Strength: strong or weak?
  - Any outlier?

CHOOSE: Simple Linear Regression Model

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

- Y : *Price*; X : *Carat*
- **Response** $Y = \beta_0 + \beta_1 X + \epsilon$ and **mean response** $\mu_Y = \beta_0 + \beta_1 X$
  β_0: **intercept**; β_1: **slope**; $\epsilon \overset{iid}{\sim} N(0, \sigma)$
- **Observed** $y = b_0 + b_1 x + e$ and **predicted** $\hat{y} = b_0 + b_1 x$.
- **Population parameters**: β_0, β_1, σ
- **Sample statistics** (estimates to the parameters): $b_0, b_1, \hat{\sigma}$
- How to find the values of b_0, b_1 and $\hat{\sigma}$?

FIT: least-squares (LS) estimation

The **least-squares regression line of Y on X** is the line that **minimizes** the sum of the squares of the **vertical distances** from the data points to the line.

In least-squares regression, we minimize

$$\sum e^2 = \sum \left(y - \hat{y}\right)^2 = \sum (y - b_0 - b_1 x)^2$$

where

$$e = y - \hat{y}$$

is defined as **residual**, the difference between the observed and the predicted response.

- LS estimation minimizes the sum of squares of residuals. Its goal is to optimize prediction - make the predicted y as close as possible to the observed y.
- It makes no assumption about the distribution of Y.

FIT: maximum likelihood estimation (MLE)

- The maximum likelihood estimation assumes that the error term in the regression model follows a Normal distribution $\epsilon \overset{iid}{\sim} N(0, \sigma)$.
- Since $Y = \mu_Y + \epsilon$ and $\epsilon \overset{iid}{\sim} N(0, \sigma)$, $Y \sim N(\mu_Y, \sigma)$, where $\mu_Y = \beta_0 + \beta_1 X$.
- It assumes Y follows a mixture of Normal distributions, with mean $\mu_Y = \beta_0 + \beta_1 X$ depending on X and SD σ that does not depend on X.

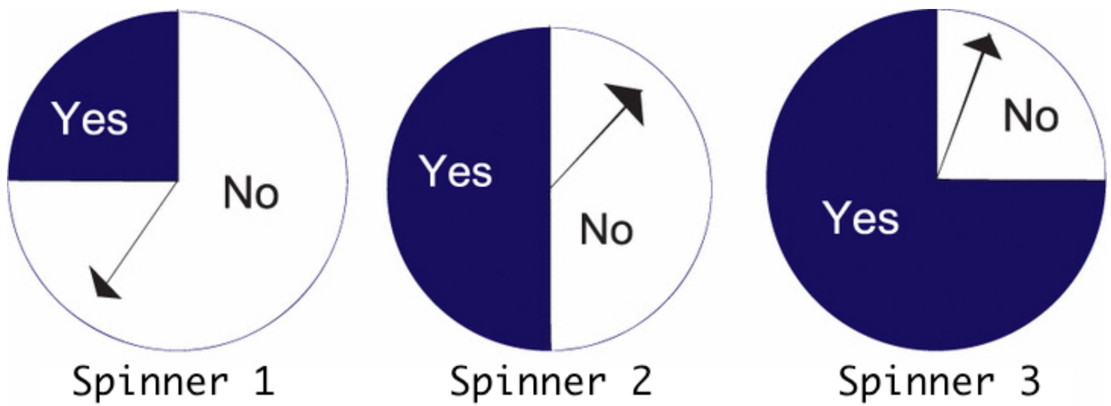
> In statistics, **maximum likelihood estimation (MLE)** is a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters.

- Goal: estimate the parameters given the observations
- How: maximize the likelihood of making the observations given the parameters
- Observations: y; parameters: β_0, β_1 and σ.

FIT: maximum likelihood estimation (MLE)

Example

A spinner is chosen, spun once, and the outcome is Yes. If you have to guess which spinner was used to get the Yes, what is your choice?



- Observation: Yes
- Parameter: which spinner
- Spinner 3. Because
  $P(Yes \mid Spinner1) = \frac{1}{4} < P(Yes \mid Spinner2) = \frac{1}{2} < P(Yes \mid Spinner3) = \frac{3}{4}$

FIT: maximum likelihood estimation (MLE)

Simple linear regression

- Observations: y
- Parameters: β_0, β_1 and σ
- We would like to estimate the parameters given the observations.
- Therefore, we will maximize P (Observations | Parameters) $= P(Y \mid \beta_0, \beta_1, \sigma)$ the likelihood of making the observations given the parameters.
- Since $Y \sim N(\beta_0 + \beta_1 X, \sigma)$, for a single observation y,

$$P(Y = y \mid \beta_0, \beta_1, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{[y-(\beta_0+\beta_1 x)]^2}{2\sigma^2}}$$

FIT: maximum likelihood estimation (MLE)

Simple linear regression

- For n observations $y_1, y_2, \cdots, y_n$,

$$P(y_1, y_2, \cdots, y_n | \beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$
$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

- We search for the values of β_0, β_1 and σ so as to maximize

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}}$$

- For estimating β_0 and β_1, this is equivalent to minimize

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

FIT: maximum likelihood estimation (MLE)

Simple linear regression

Therefore, the least squares method and the maximum likelihood method result in the same estimates for β_0, β_1 and σ.

$$\text{Slope } b_1 = r\frac{s_y}{s_x}, \text{ Intercept } b_0 = \bar{y} - b_1\bar{x},$$

$$\text{Residual standard error } \hat{\sigma} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} = \sqrt{\frac{\sum(y-b_0-b_1x)^2}{n-2}}.$$

- $\bar{x}, \bar{y}$: mean of x and y; s_x, s_y: SD of x and y; r: correlation of x and y

Note: this does not mean that the LS method is equivalent to the MLE method. In terms of estimation, they get the same results in simple linear regression. But the MLE method makes Normal assumption about the data, which facilitates model inferences.

FIT: Simple linear regression model in R

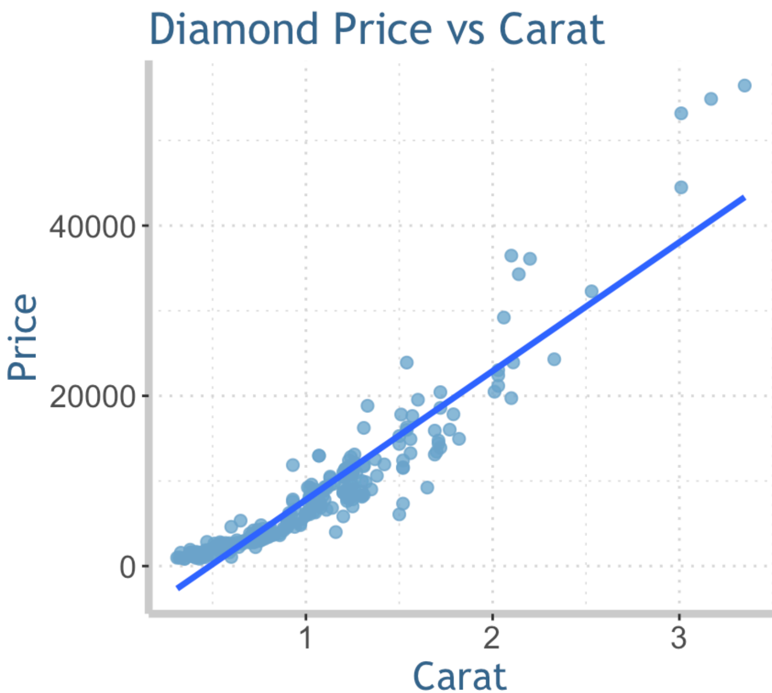
```r
summary(diaSLR <- lm(Price ~ Carat, data=Diamonds))
```

```
## Call:
## lm(formula = Price ~ Carat, data = Diamonds)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9278.5 -1341.7  -236.2  1230.9 14991.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7341.7      361.1  -20.33   <2e-16 ***
## Carat        15130.1      331.0   45.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2860 on 305 degrees of freedom
## Multiple R-squared:  0.8726, Adjusted R-squared:  0.8722
## F-statistic:  2090 on 1 and 305 DF,  p-value: < 2.2e-16
```

- $b_0 = -7341.7$
- $b_1 = 15130.1$
- Estimated regression line $\hat{y} = -7341.7 + 15130.1x$
- $\hat{\sigma} = 2860$ on $n - 2 = 305$ degrees of freedom

FIT - Regression line and scatterplot

```
ggplot(data=Diamonds, aes(x=Carat, y=Price))+
  geom_point(color="skyblue3", size=2, alpha=0.8)+ # Scatterplot
  geom_smooth(method='lm', size=1.2, se=F)+ # Add the regresion line
  ggtitle("Diamond Price vs Carat")
```



$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X &+ \quad \epsilon \\
y &= b_0 + b_1 x &+ \quad e
\end{aligned}
$$

The **estimated regression line**

$$\hat{y} = b_0 + b_1 x = -7341.7 + 15130.1x$$

- $b_0 = -7341.7$: when *Carat* is 0, *Price* is -7341.7 (value of b_0 is sometimes practically not meaningful).
- $b_1 = 15130.1$: as *Carat* increases by 1 unit, *Price* increases \$15130.1.

ASSESS model

- SLR model assumes that Y depends on X

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- A simpler model would be that Y does NOT depend on X, which means that $\beta_1 = 0$,

$$Y = \beta_0 + \epsilon$$

- Therefore, the hypotheses of a simple linear regression model are
  $H_0 : \beta_1 = 0$, there is no linear relationship between Y and X;
  $H_a : \beta_1 \neq 0$, there is a linear relationship between Y and X.

ASSESS model: inference for intercept & slope

To test whether the population slope β_1 is different from zero, the hypotheses are $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$, and the test statistic is

$$t = \frac{b_1}{SE_{b_1}} \sim t(n-2).$$

If the proper conditions hold, we compute the P-value from the $t(n-2)$ distribution.

The level C confidence intervals for β_0 and β_1 are

$$b_0 \pm t^* SE_{b_0}, \quad b_1 \pm t^* SE_{b_1}$$

where t^* is the critical value for the $t(n-2)$ density curve to obtain the desired confidence level C.

ASSESS model: inference for intercept & slope

```
summary(diaSLR)$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -7341.7      361.1  -20.33   <2e-16 ***
## Carat        15130.1      331.0   45.72   <2e-16 ***
```

t test for the slope β_1

- $b_1 = 15130.1, SE_{b_1} = 331.0$
- $t = \frac{b_1}{SE_{b_1}} = 45.7 \sim t(305)$
- $P < 2 \times 10^{-16} << 0.05$
- We reject H_0 that $\beta_1 = 0$. The linear relationship between number of carats of diamonds and price is highly significant at level 0.05.
- Note: usually we do not test the intercept. But R always provides a t test for the intercept with $H_0 : \beta_0 = 0$, which often does not have practical meaning.

ASSESS model: inference for intercept & slope

```r
# Get the 95% confidence intervals for the intercept and slope
confint(diaSLR)
```

```
##                  2.5 %    97.5 %
## (Intercept) -8052.184 -6631.239
## Carat       14478.881 15781.404
```

- **95% confidence interval for the intercept β_0**: $[-8052.2, -6631.2]$

  We are 95% confident (about the method) that the interval $[-8052.2, -6631.2]$ will contain the true population intercept.

- **95% confidence interval for the slope β_1**: $[14478.9, 15781.4]$

  We are 95% confident (about the method) that the interval $[14478.9, 15781.4]$ will contain the true population slope.
  This interval does not contain 0 $\Leftrightarrow$ the t test for the slope is significant.

ASSESS error: model assumptions

- **Linearity**: there is a linear relationship between Y and X.

$$\epsilon \overset{iid}{\sim} N(0, \sigma)$$

- **Zero mean**: mean of the errors is 0.
- **Constant variance**: the variability in the errors is the same for all values of the explanatory variable.
- **Normality**: the errors follow a normal distribution (to use the t distribution for inference).
- **Independence and randomness**: the errors are independent from one another and the data are obtained randomly.

ASSESS error: model assumptions

Linearity

- Scatterplot of response y on explanatory x
- Scatterplot of residuals e on fitted values $\hat{y}$

Zero mean - always true

Constant variance

- Scatterplot of residuals e on fitted values $\hat{y}$
- Breusch-Pagan test for H_0: constant variance

Normality

- Normal Q-Q plot of the residuals (sometimes histogram of residuals is helpful)

Independence and randomness - check data collecting process

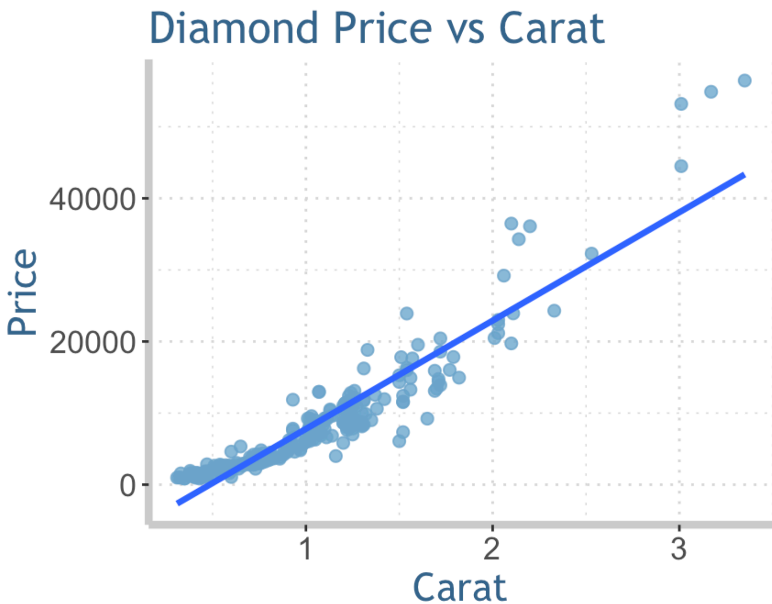
ASSESS error: R codes

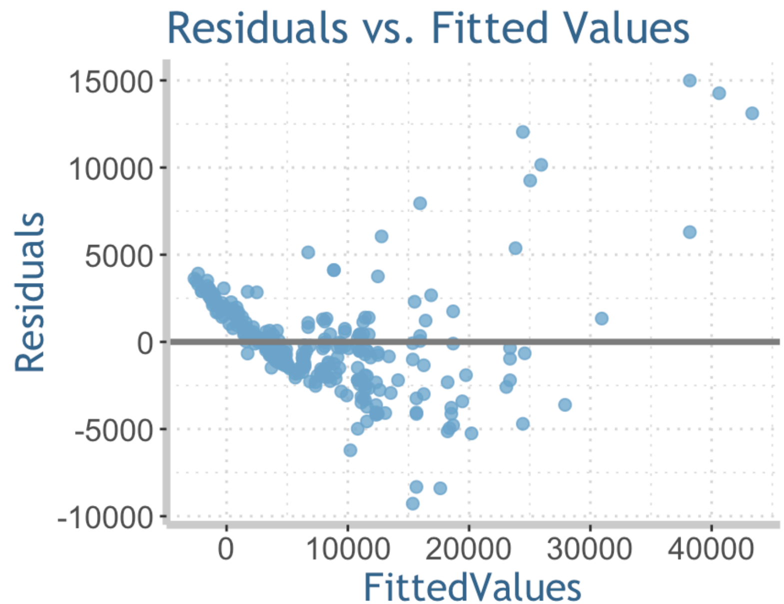
```r
# Scatterplot
ggplot(data=Diamonds, aes(x=Carat, y=Price))+
  geom_point(color="skyblue3", size=2, alpha=0.8)+ # Scatterplot
  geom_smooth(method='lm', size=1.2, se=F)+ # Add the regresion line
  ggtitle("Diamond Price vs Carat")

Assess <- data.frame(Residuals=diaSLR$residuals,
                     FittedValues=diaSLR$fitted.values)
# Residuals vs. Fitted Values
ggplot(data=Assess, aes(x=FittedValues, y=Residuals))+
  geom_point(color="skyblue3", size=2, alpha=0.8)+
  geom_hline(yintercept=0, size=1.2, colour="grey50")+ # Add y=0 line
  ggtitle("Residuals vs. Fitted Values")

# Normal Q-Q plot
ggplot(data=Assess, aes(sample = scale(Residuals)))+
  stat_qq(size=3, color="skyblue3", alpha=0.8)+
  geom_abline(intercept=0, slope=1, size=1.2, colour="grey50")+ # Add y=x line
  ggtitle("Normal Q-Q Plot")
```

ASSESS error: Linearity



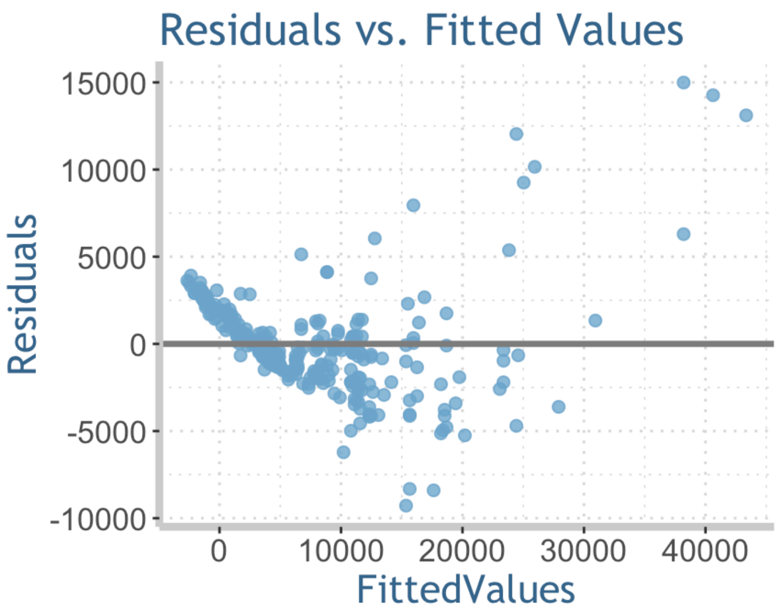


- **Scatterplot**: linear or curved?
- **Residuals vs. fitted values plot**: any pattern? If the relationship between y and x is linear, then this plot of e vs. $\hat{y}$ should have no pattern. If there is any pattern, possibly relationship between the two variables is non-linear.

ASSESS error: Constant variance

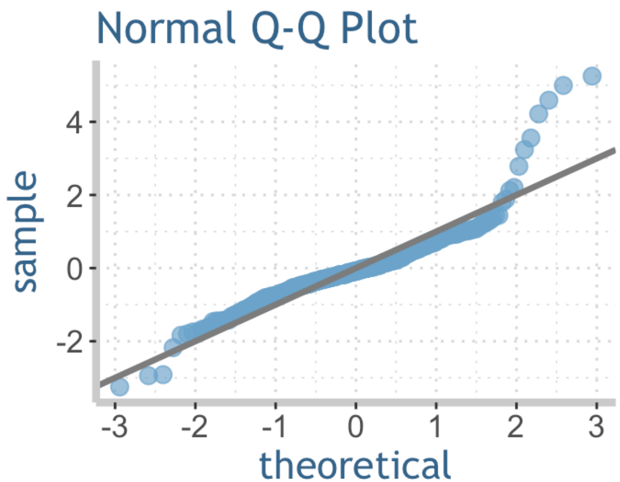
```r
library(lmtest)
bptest(diaSLR) # BP test
```

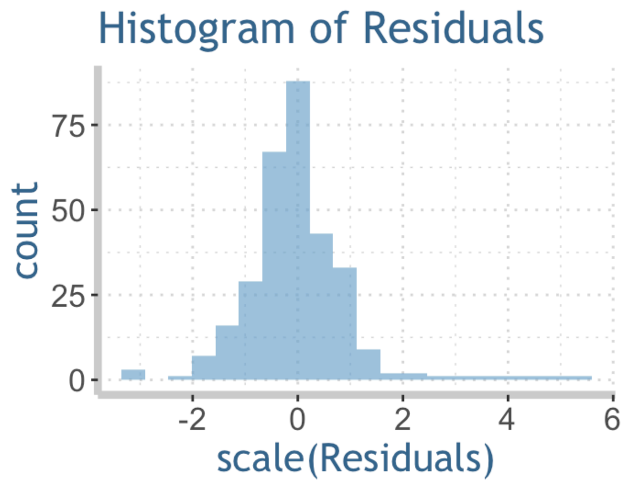
```
##
##  studentized Breusch-Pagan test
##
## data:  diaSLR
## BP = 102.3, df = 1, p-value < 2.2e-16
```



- **BP test**. H_0: the variance of the residuals is a constant and does not depend on the explanatory variable.
  - If $P \leq 0.05$, we reject H_0 thus the constant variance assumption is violated.
  - If $P > 0.05$, we cannot reject H_0 thus the constant variance assumption is satisfied.
- **Residuals vs. Fitted Values**. Is the spread of the residuals roughly the same for different fitted values?

ASSESS error: Normality





- **Normal Q-Q plot**: all the points lie close to the $y = x$ line?

  Most points lie close to the line; histogram is symmetric. But both plots have quite heavy tails on the right.
- **Conclusion**. The residuals vs. fitted values plot shows a clear pattern; the spread of the residuals are not roughly the same for different fitted values; BP test shows significance. Therefore, the linearity and constant variance are strongly violated. Normality assumption might be slightly violated.

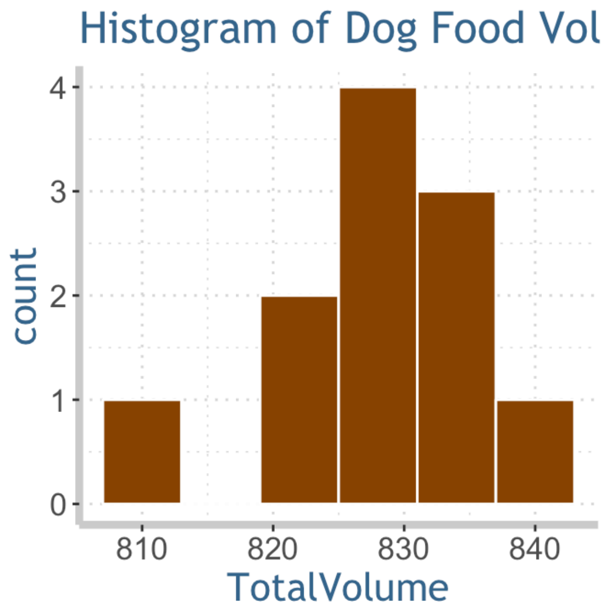
U.K. Pets

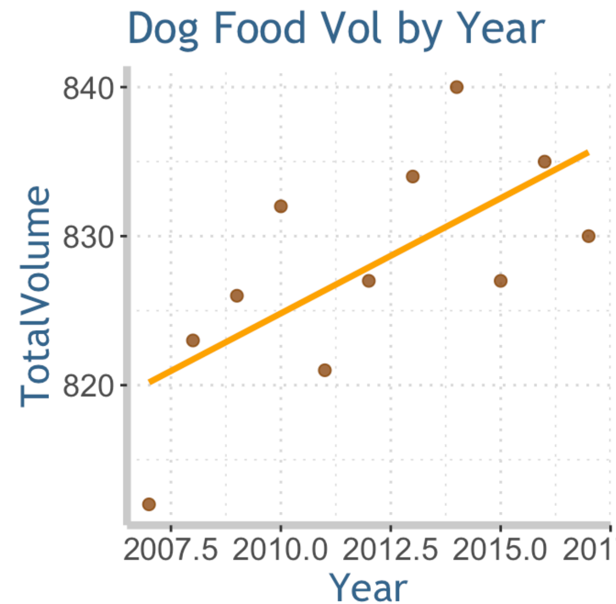
- Graph available at the website.
- Data source: Pet Food Manufacturers Association
- PFMA publishes pet population and food market data every year.

U.K. dog food volume

```
dogfood[, c("Year", "TotalVolume")] # Total volume in 1,000 tons
```

```
##    Year TotalVolume
## 1  2007         812
## 2  2008         823
## 3  2009         826
## 4  2010         832
## 5  2011         821
## 6  2012         827
## 7  2013         834
## 8  2014         840
## 9  2015         827
## 10 2016         835
## 11 2017         830
```





- Response Y: *TotalVolume*, mean 827.9, sd 7.6; Explanatory X: *Year*, 2007 ~ 2017.
- Correlation coefficient: 0.67.
- Model: $Y = \beta_0 + \beta_1 X + \epsilon$, where $\epsilon \overset{iid}{\sim} N(0, \sigma)$.

U.K. dog food volume

```r
dogSLR <- lm(TotalVolume ~ Year, data=dogfood)
summary(dogSLR)$coefficient
```

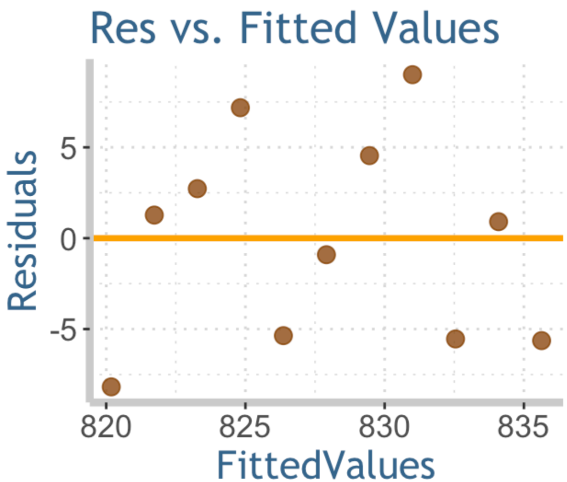
```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2281.5455  1147.7873  -1.988   0.0781 .
## Year            1.5455     0.5705   2.709   0.0240 *
```

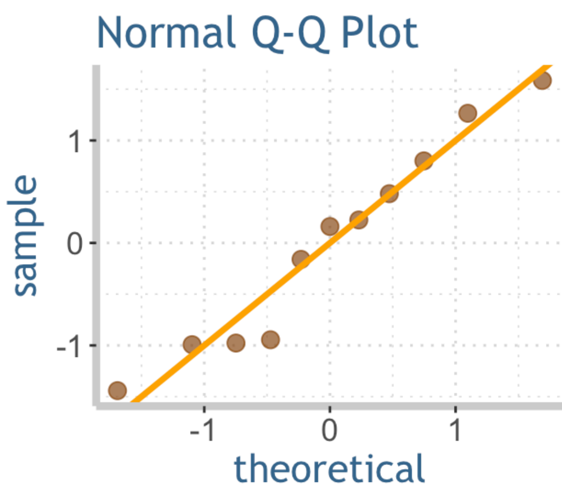
```r
confint(dogSLR)
```

```
##                      2.5 %     97.5 %
## (Intercept) -4878.0208030 314.929894
## Year            0.2549614   2.835948
```

- Estimated regression line: $\hat{y} = -2281.5 + 1.5x$
  - U.K. dog food volume increases 1500 tons every year.
- t test for the slope: $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$. $t = 2.7$ and $P = 0.024 < 0.05$. There is a statistical significant linear relationship between *TotalVolume* and *Year*.
- 95% CIs for β_0 and β_1: $[-4878.0, 314.9]$ and $[0.255, 2.836]$.

U.K. dog food volume






```
##
##  studentized Breusch-Pagan test
##
## data:  dogSLR
## BP = 0.045207, df = 1, p-value = 0.8316
```

Checking assumptions:

- Scatterplot shows a linear trend.
- Residuals vs. fitted values plot has no clear pattern and the spread of the points is roughly the same for fitted values and symmetric about the $y = 0$ line.
- BP test has $BP = 0.045$ and $P = 0.832 > 0.05$.
- Points on the Normal Q-Q plot all lie very close to the $y = x$ line.
- No evidence of violation in the linearity, constant variance or Normality assumption is found.

Summary

CHOOSE

- Exploratory data analysis; Model definition $Y = \beta_0 + \beta_1 X + \epsilon$ 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