STAT021 Statistical Methods II

Lecture 10 SLR ANOVA and Transformation

Lu Chen
Swarthmore College
10/4/2018

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

Outline

- Simple linear regression ANOVA
  - Sum of squares and degree of freedom
  - 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

Simple linear regression ANOVA

$$
\begin{array}{rcccccl}
 & \text{Data} & = & \text{Model} & + & \text{Error} & \\
\text{ANOVA:} & Y & = & \mu + \alpha_k & + & \epsilon, & \text{where } k = 1, 2, \cdots, K \text{ and } \epsilon \overset{iid}{\sim} N(0, \sigma \\
 & y & = & \bar{y} + \bar{y}_k - \bar{y} & + & y - \bar{y}_k & \\
 & y - \bar{y} & = & \bar{y}_k - \bar{y} & + & y - \bar{y}_k & \\
 & & & & & & \\
\text{SLR:} & Y & = & \beta_0 + \beta_1 X & + & \epsilon, & \text{where } \epsilon \overset{iid}{\sim} N(0, \sigma) \\
 & y & = & b_0 + b_1 x & + & e & \\
 & y & = & \hat{y} & + & y - \hat{y} & \\
 & y - \bar{y} & = & \hat{y} - \bar{y} & + & y - \hat{y} &
\end{array}
$$

- In simple linear regression, $\hat{y} = b_0 + b_1 x$

Sum of squares and degree of freedom

Sum of squares:	$SSTotal$	$=$	$SSModel$	$+$	SSE
	$\sum(y-\bar{y})^2$	$=$	$\sum(\hat{y}-\bar{y})^2$	$+$	$\sum(y-\hat{y})^2$
	Total variability in response Y	$=$	Variability explained by the SLR model	$+$	Variability in residuals
Degrees of freedom:	df_{Total}	$=$	df_{Model}	$+$	df_{Error}
	$n-1$	$=$	1	$+$	$n-2$

- $\hat{y} = b_0 + b_1 x$, which involves two statistics b_0 and b_1. Therefore, the degree of freedom for the *Model* term is $2-1=1$ and for the *Error* term $n-2$.

Mean square, F test and R-squared

$$MSModel = \frac{SSModel}{df_{Model}} = \frac{\sum(\hat{y}-\bar{y})^2}{1}, \quad MSE = \frac{SSE}{df_{Error}} = \frac{\sum(y-\hat{y})^2}{n-2}$$

$$F = \frac{MSModel}{MSE} = \frac{\frac{\sum(\hat{y}-\bar{y})^2}{1}}{\frac{\sum(y-\hat{y})^2}{n-2}} \sim F(1, n-2)$$

$$R^2 = \frac{SSModel}{SSTotal} = \frac{\sum(\hat{y}-\bar{y})^2}{\sum(y-\bar{y})^2}$$

- MSE is the estimate to the variance of error. The residual standard error is

$$\hat{\sigma} = \sqrt{MSE} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}$$

- F test indicates the significance of the SLR model. R^2 measures the strength of (the fraction of variability explained by) the SLR model.

SLR ANOVA table

To test the effectiveness of the simple linear model, the hypotheses are

$H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$.

The **ANOVA table** is

	Degree of Freedom	Sum of Squares	Mean Square	F statistic	P-value
Model	1	$SSModel$	$MSModel$	$F = \frac{MSModel}{MSE}$	$P(F_{1,n-2} > F)$
Error	$n-2$	SSE	MSE		
Total	$n-1$	SST			

If the conditions for the simple linear regression model hold, the P-value is obtained from the upper tail of an F-distribution with 1 and $n-2$ degrees of freedom.

SLR ANOVA 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
##
## 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
```

- *Multiple R-squared* $R^2 = 0.8726$
- $F = 2090$
- Degrees of freedom: 1 and $n - 2 = 305$
- $P < 2.2 \times 10^{-16}$
- *Adjusted R-squared* will be discussed in multiple linear regression.

SLR ANOVA in R

```
anova(diaSLR) # obtain the ANOVA table for the SLR model
```

```
## Analysis of Variance Table
##
## Response: Price
##            Df     Sum Sq    Mean Sq F value    Pr(>F)
## Carat       1 1.7091e+10 1.7091e+10  2089.9 < 2.2e-16 ***
## Residuals 305 2.4943e+09 8.1779e+06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We reject the null hypohtesis that $\beta_1 = 0$ at level 0.05. The *Carat* variable in the linear regression model has a significant effect in explaining the response variable $Price$.
- $R^2 = 0.8726$. About 87% of the variability in $Price$ is explained by the SLR model that involves *Carat*.

Regression and correlation

- Parameters of a simple linear regression model: β_0, β_1 and σ. Their estimates are

$$b_1 = \boldsymbol{r}\frac{s_y}{s_x}, \quad b_0 = \bar{y} - b_1\bar{x}, \quad \hat{\sigma} = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}}.$$

- $\boldsymbol{r}$: correlation coefficient
  - It is an estimate to the population correlation ρ.
  - It measures of the strength of the linear association between two quantitative variables.
  - $-1 \leq r \leq 1$; $r = 0$ means no linear association.
- Correlation coefficient r is related to the regression slope b_1

  $r = 0 \iff b_1 = 0$. Testing $\beta_1 = 0$ is equivalent to testing $\rho = 0$.
- Correlation coefficient r is also related to the regression R^2, $R^2 = r^2$.

Regression and correlation

```r
cor(Diamonds$Price, Diamonds$Carat)
```

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

```r
cor(Diamonds$Carat, Diamonds$Price)
```

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

```r
lm(Price ~ Carat, data=Diamonds)$coefficients
```

```
## (Intercept)       Carat
##   -7341.712   15130.142
```

```r
lm(Carat ~ Price, data=Diamonds)$coefficients
```

```
##  (Intercept)        Price
## 5.473680e-01 5.767599e-05
```

```r
summary(diaSLR)$r.squared; 0.9341552^2
```

```
## [1] 0.8726459
```

```
## [1] 0.8726459
```

- $r = 0.934$, strong positive correlation.
- Correlation of Y and X is the same as correlation of X and Y.
- Regression of Y on X is different from regression of X on Y
- In SLR, ANOVA R^2 is exactly correlation squared.

t test for correlation

Let ρ denote the population correlation, the hypotheses are

$H_0 : \rho = 0$ and $H_a : \rho \neq 0$

and the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t(n-2)$$

If the conditions for the simple linear model hold, we find the P-value using the t-distribution with $n-2$ degrees of freedom.

t test for correlation

```
cor.test(~ Price + Carat, data=Diamonds)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Price and Carat
## t = 45.715, df = 305, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.9182342 0.9470616
## sample estimates:
##       cor
## 0.9341552
```

- $t = 45.72$, $df = n - 2 = 305$, $P < 2.2 \times 10^{-16} < 0.05$.
- We reject H_0 that $\rho = 0$ at level 0.05. There is a highly significant linear association between *Price* and *Carat*.
- 95% C.I.: [0.918, 0.947]

Three tests for linear relationship?

Response variable: *Price*; Explanatory variable: *Carat*

	t test for slope	F test for model	t test for correlation
H_0	$\beta_1 = 0$	$\beta_1 = 0$	$\rho = 0$
Test statistic	$t = 45.72$	$F = 2090$	$t = 45.72$
Distribution	$t(n-2) = t(305)$	$F(1, n-2) = F(1, 305)$	$t(n-2) = t(305)$
P**-value**	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$	$< 2.2 \times 10^{-16}$

- $\beta_1 = 0 \iff \rho = 0$
- $F = 2090 = t^2 = 45.72^2$
- If $t \sim t(df)$, $F = t^2 \sim F(1, df)$
- The three tests are equivalent in the simple linear regression setting.

Three tests for linear relationship?

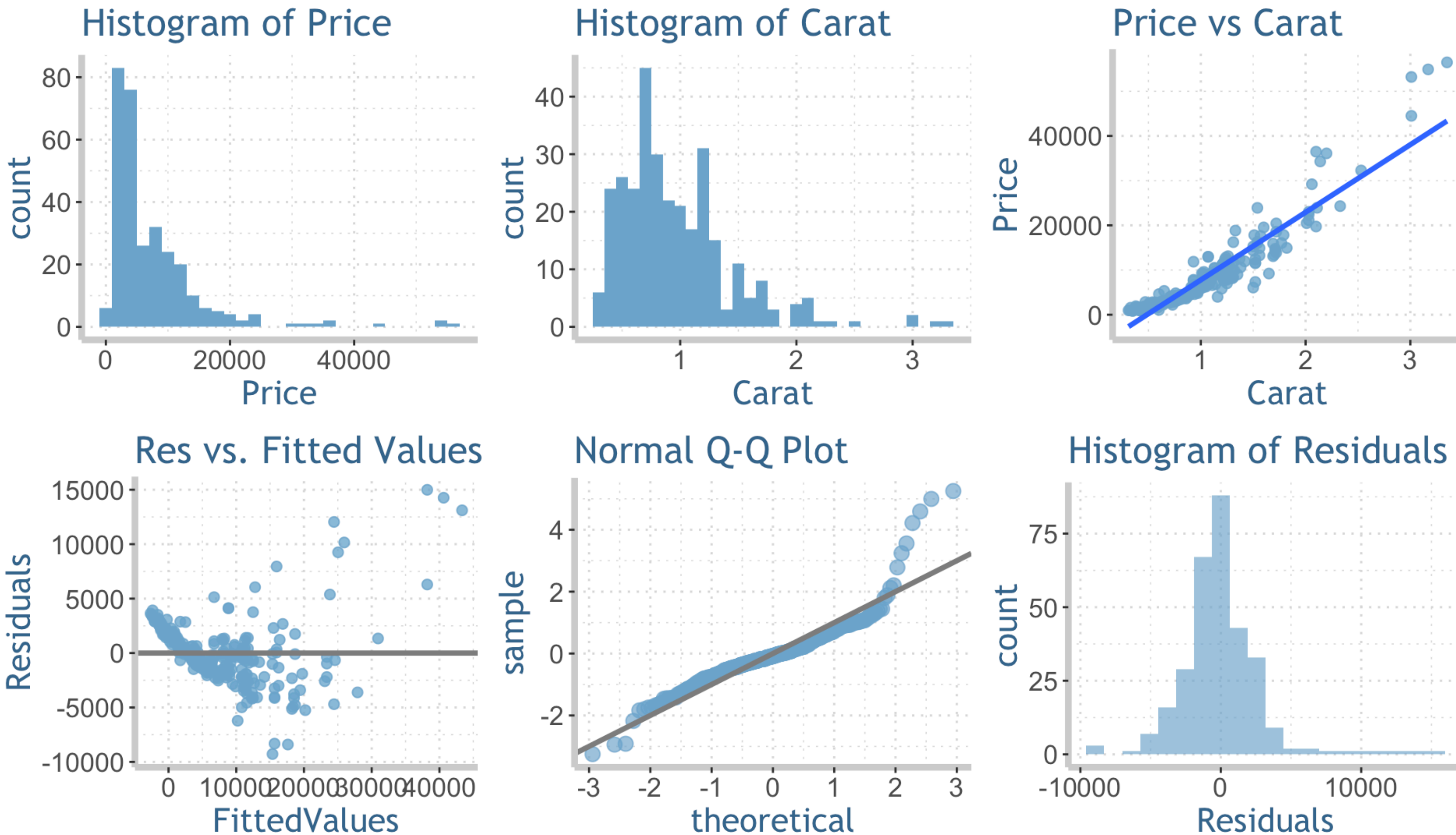
Why do we need three equivalent tests for a linear relationship?

- While the results are equivalent in the simple linear regression case, we will see that these tests take on different roles in multiple linear regression model.

In multiple linear regression

- t test for the slope: relationship between the response and the explanatory considering other explanatory variables are in the model.
- ANOVA F test for the model: relationship between the response and all the explanatory variables ($H_0 : \beta_1 = \beta_2 = \cdots = \beta_K = 0$).
- t test for the correlation: relationship between the response and the explanatory without considering other explanatory variables.

Transformation: Diamond price



Transformation: Diamond price

- BP test for constant variance is highly significant ($BP = 102, P < 2.2 \times 10^{-16}$).
- The linearity and constant variance assumptions are strongly violated; the Normality assumption may be slightly violated.
- Since the distribution of *Price* and the residuals are skewed to the right, let's try natural logarithm transformation.
- Denote *log(Price)* as Y,

$$Y = \beta_0 + \beta_1 X + \epsilon, \text{ where } \epsilon \overset{iid}{\sim} N(0, \sigma)$$

- Note: in the new model, the relationship being evaluated is between *log(Price)* and *Carat*.

Transformation: log(Price) ~ Carat

```r
diaSLR2 <- lm(log(Price) ~ Carat, data=Diamonds)
summary(diaSLR2)$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.91729    0.04174  165.72   <2e-16 ***
## Carat        1.63724    0.03826   42.79   <2e-16 ***
```

```r
summary(diaSLR2)$r.squared
```

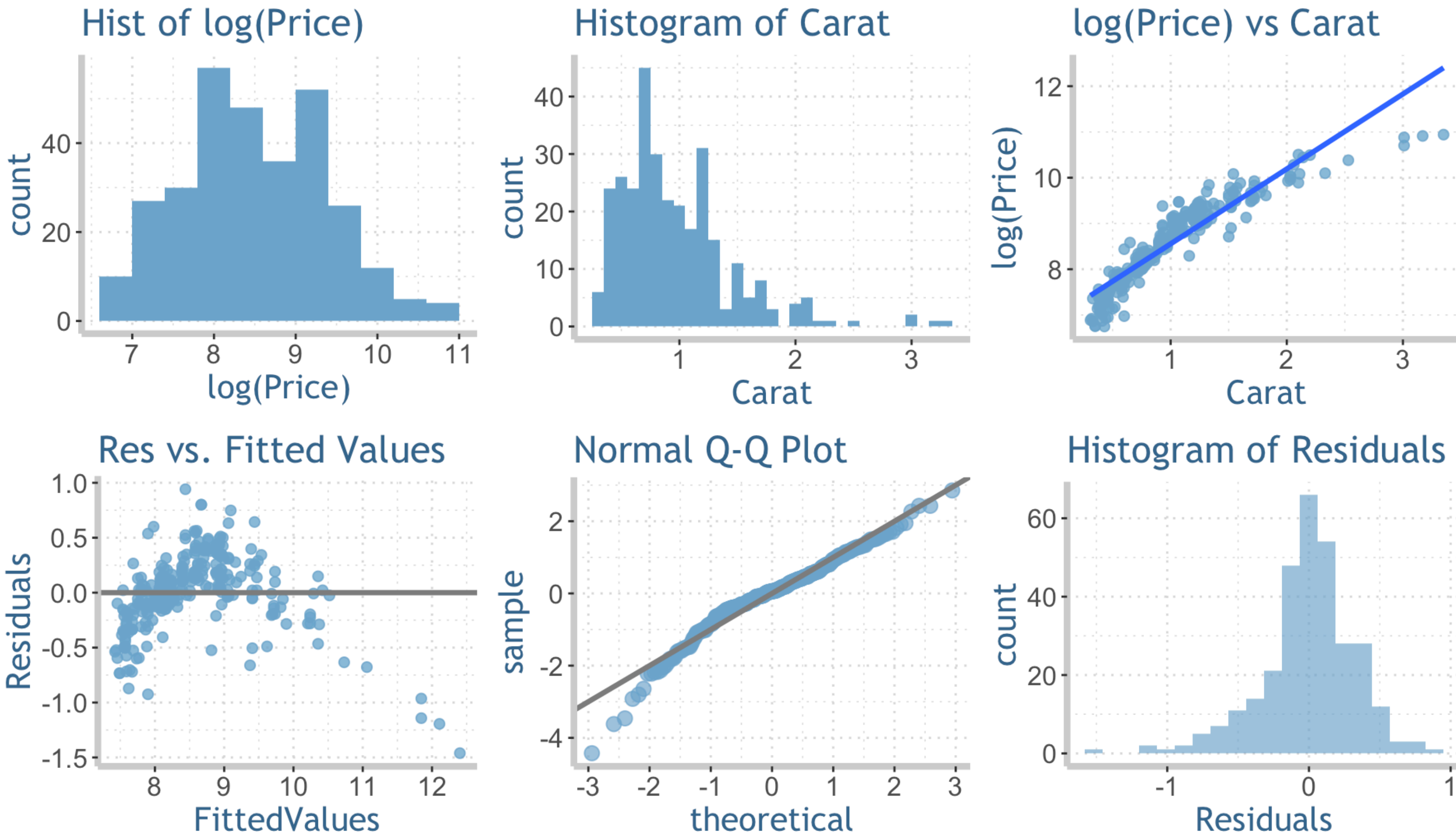
```
## [1] 0.8572019
```

```r
library(lmtest); bptest(diaSLR2) # BP test
```

```
##
##  studentized Breusch-Pagan test
##
## data:  diaSLR2
## BP = 39.487, df = 1, p-value = 3.303e-10
```

- $\widehat{log(Price)} = 6.9 + 1.6 \times Carat$
- $t = 42.9, P << 0.05$
- $R^2 = 0.86$
- $BP = 39.5, P = 3.3 \times 10^{-10}$

Transformation: log(Price) ~ Carat



Transformation: log(Price) ~ log(Carat)

```r
# Let's also transform the explanatory variable Carat
diaSLR3 <- lm(log(Price) ~ log(Carat), data=Diamonds)
summary(diaSLR3)$coefficients
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.76116    0.01311  668.49   <2e-16 ***
## log(Carat)   1.79331    0.02669   67.18   <2e-16 ***
```

```r
summary(diaSLR3)$r.squared
```

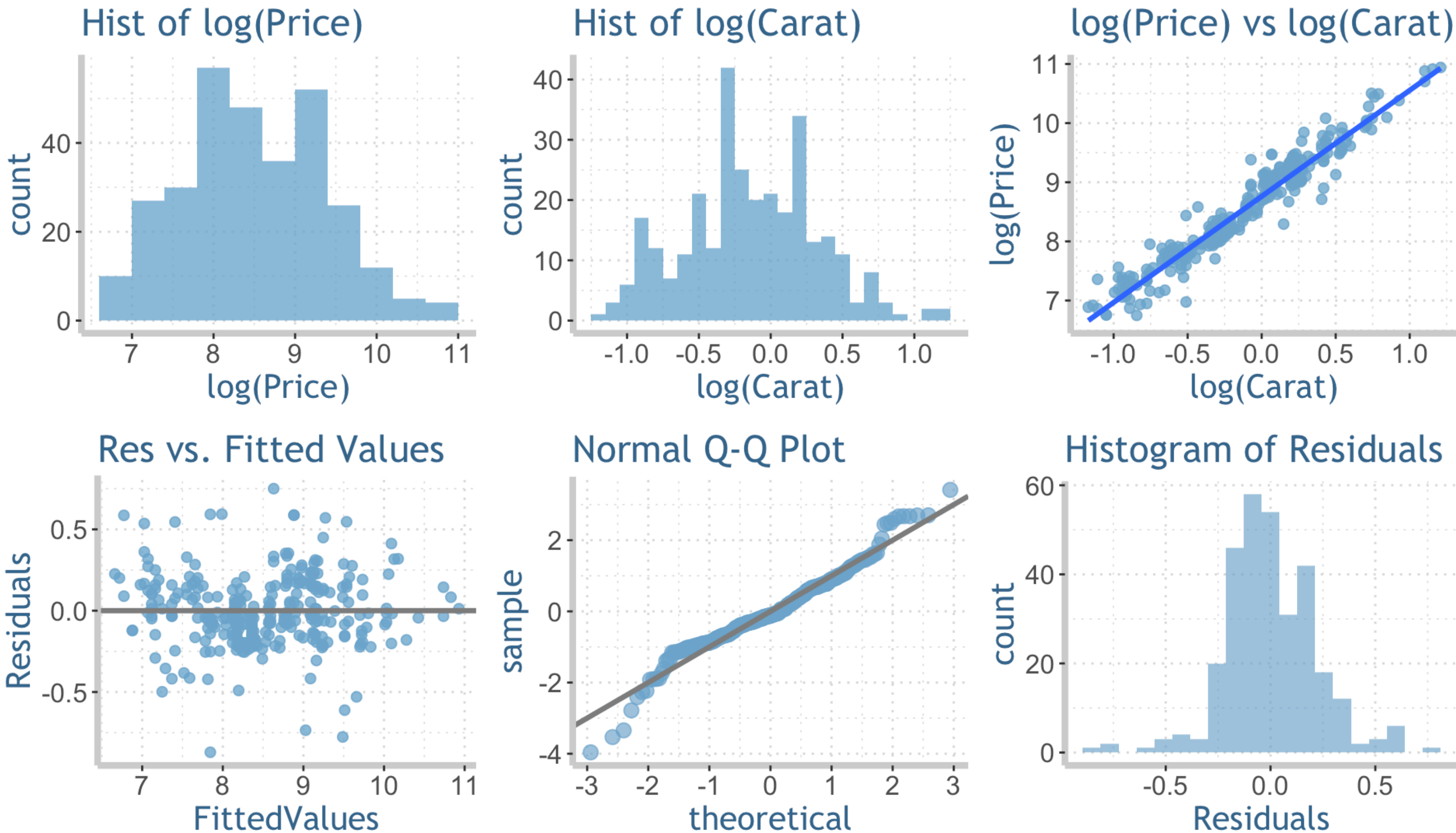
```
## [1] 0.9366962
```

```r
bptest(diaSLR3) # BP test
```

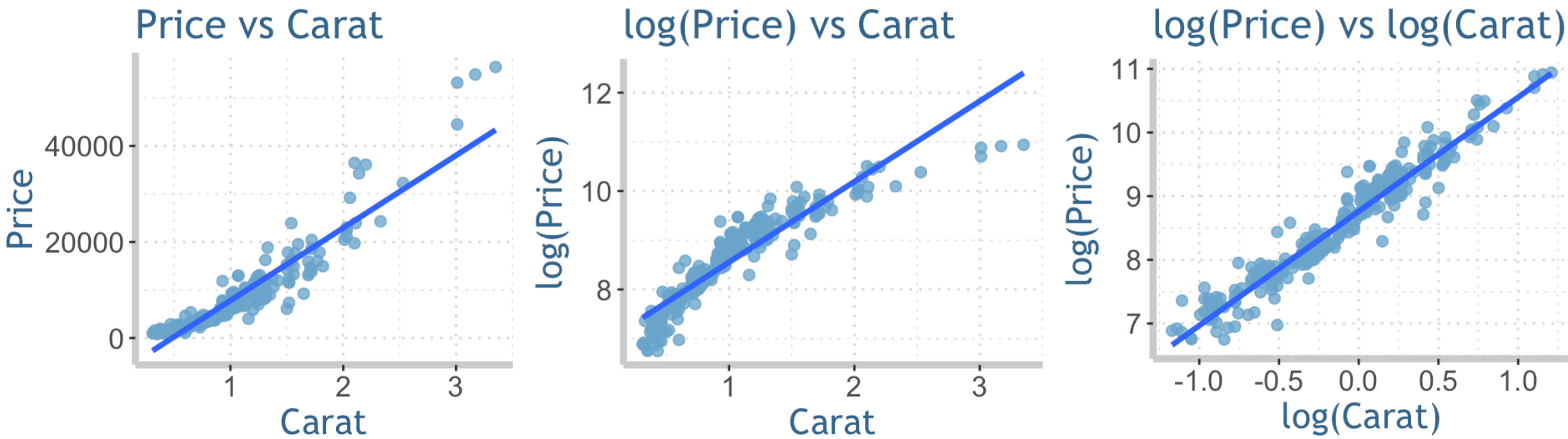
```
##
##  studentized Breusch-Pagan test
##
## data:  diaSLR3
## BP = 0.010945, df = 1, p-value = 0.9167
```

- $\widehat{log(Price)} = 8.8 + 1.8 \times log(Carat)$
- $t = 67.2, P << 0.05$
- $R^2 = 0.94$
- $BP = 0.01, P = 0.912 > 0.05$

Transformation: log(Price) ~ log(Carat)

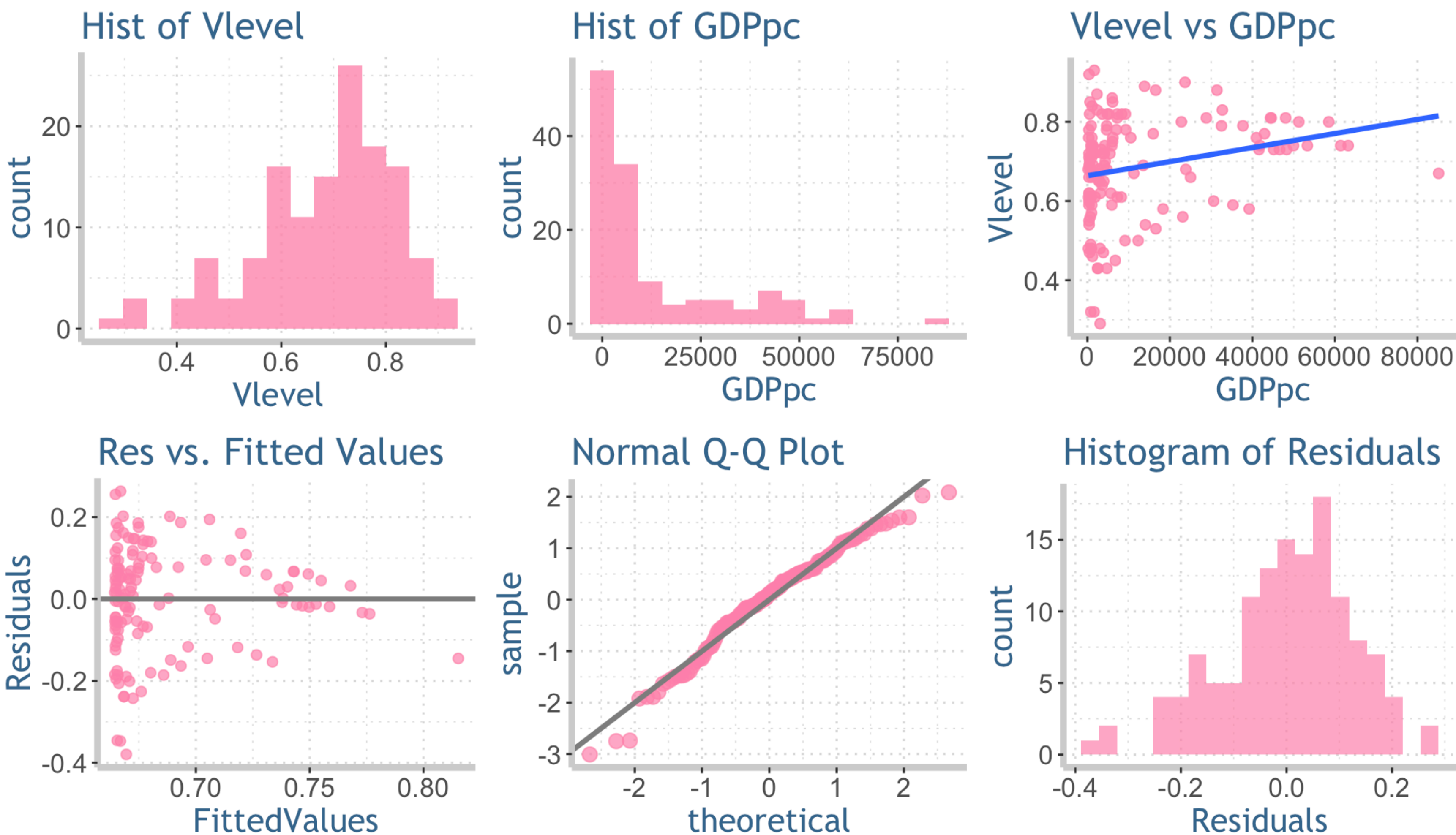


Best model: log(Price) ~ log(Carat)

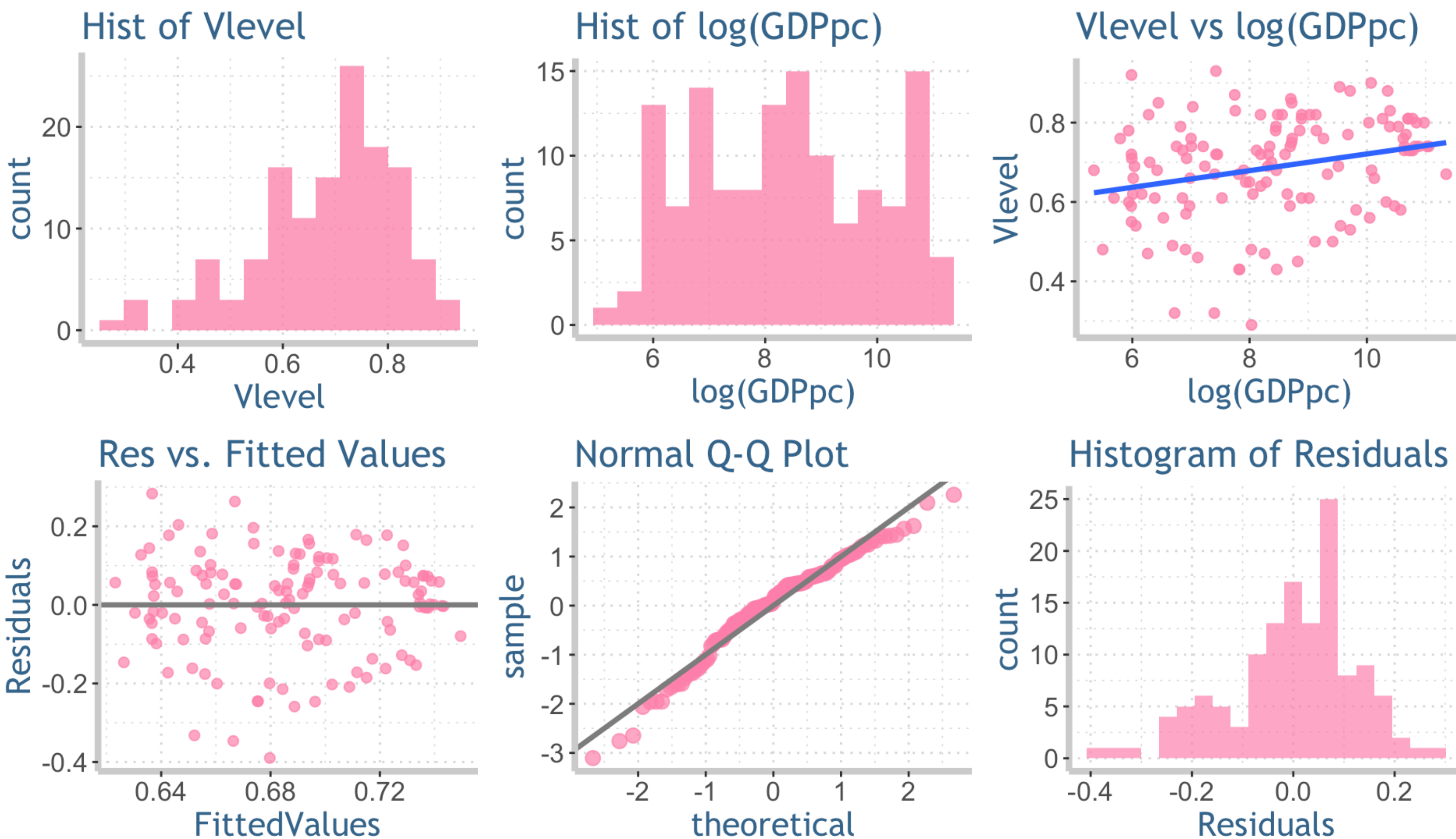


	Price ~ Carat	log(Price) ~ Carat	log(Price) ~ log(Carat)
t **test for slope**	$t = 45.7$, tiny P	$t = 42.9$, tiny P	$t = 67.2$, tiny P
R^2	0.873	0.857	0.937
BP test	$BP = 102$, tiny P	$BP = 39$, tiny P	$BP = 0.01$, $P = 0.917$

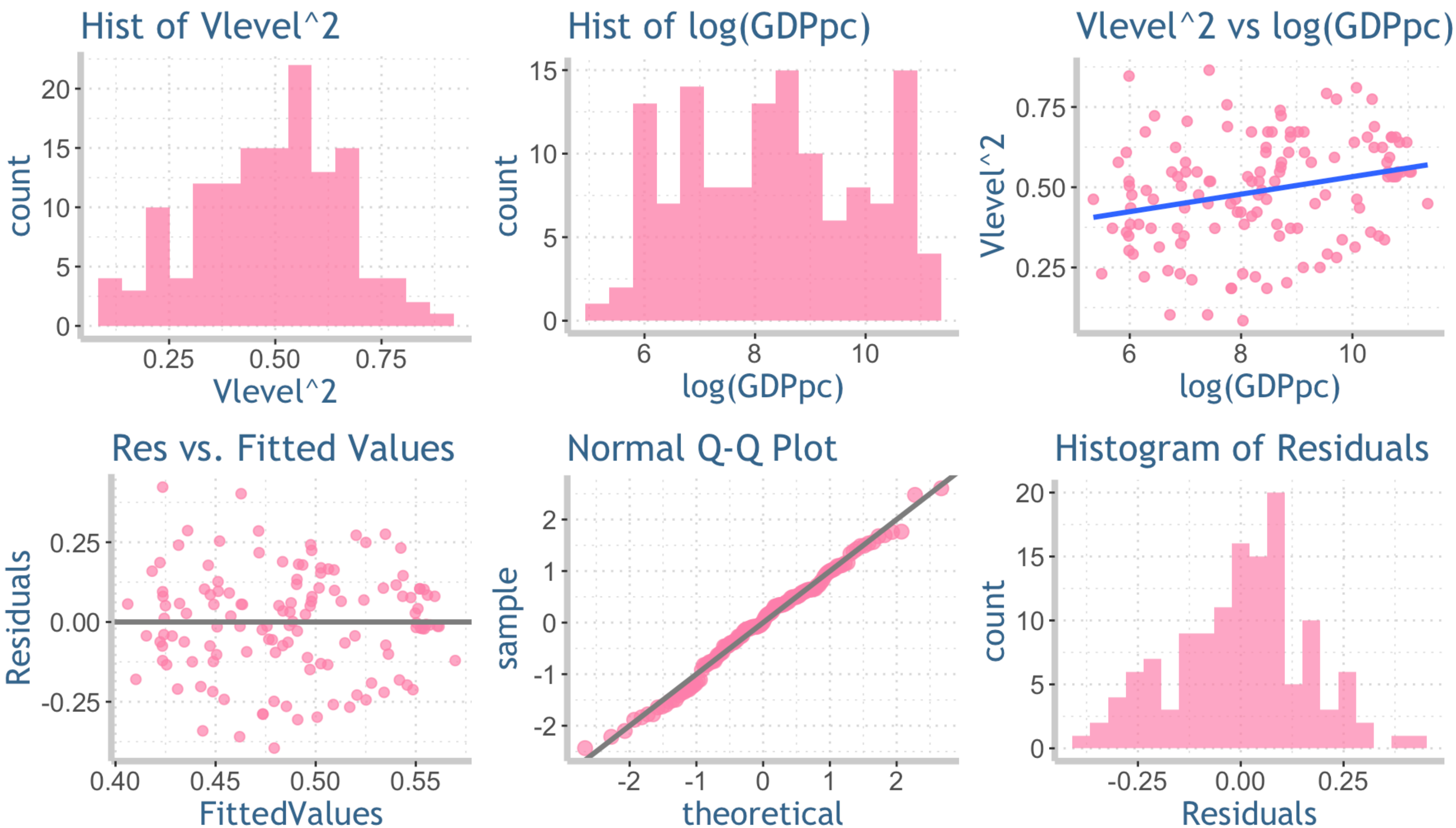
Transformation: Valentine's Day love level



Transformation: Vlevel ~ log(GDPpc)



Transformation: Vlevel^2 ~ log(GDPpc)



Best model: Vlevel^2 ~ log(GDPpc)

	Vlevel ~ GDPpc	**Vlevel ~ log(GDPpc)**	**Vlevel^2 ~ log(GDPpc)**
t **test for slope**	$t = 2.8, P = .0055$	$t = 3.1, P = .0026$	$t = 3.1, P = .0026$
R^2	0.058	0.068	0.068
BP test	$BP = 5.0, P = .025$	$BP = 2.4, P = .121$	$BP = 1.7, P = 0.194$

- The second and the third model are very similar.
- The third one is slightly better in terms of the t test P value and R^2 .
- The third model is the best while the second one is also very good.

Some notes

- Recall the ANOVA model for *Vlevel* and *GDPpc*, where the latter is categorized as a categorical variable with 4 categories, *VeryLow*, *Low*, *Medium* and *High*.
- That ANOVA model has $R^2 = 0.087$.
- Our SLR model for $Vlevel^2$ versus $log(GDPpc)$ has $R^2 = 0.068$.
- Categorization usually causes loss of information. But why the SLR model explains even less variability than the ANOVA model?
- Linear relationship between any two variables is in fact a very strong assumption while categorization allows more flexibility.
- There is NO perfect model. There is no guarantee that transformation will eliminate all problems.
- Consider the following when you choose a model: are model assumptions violated? How significant is the t or F test for $\beta_1 = 0$? Is R^2 large?

Summary

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