STAT011 Statistical Methods I

Lecture 24 Simple Linear Regression III

Lu Chen
Swarthmore College
4/25/2019

Review

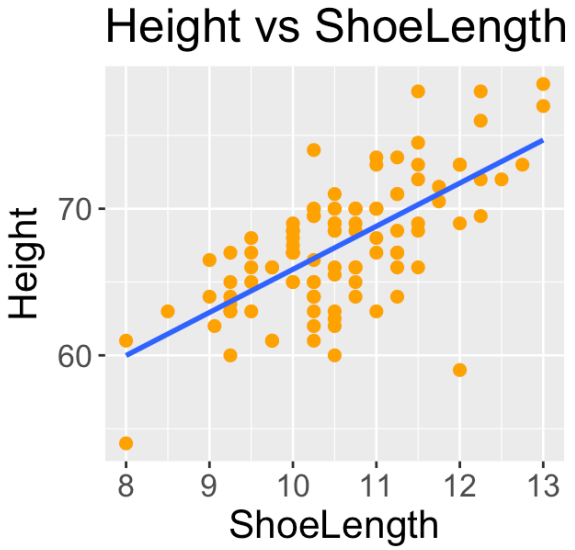
Simple linear regression

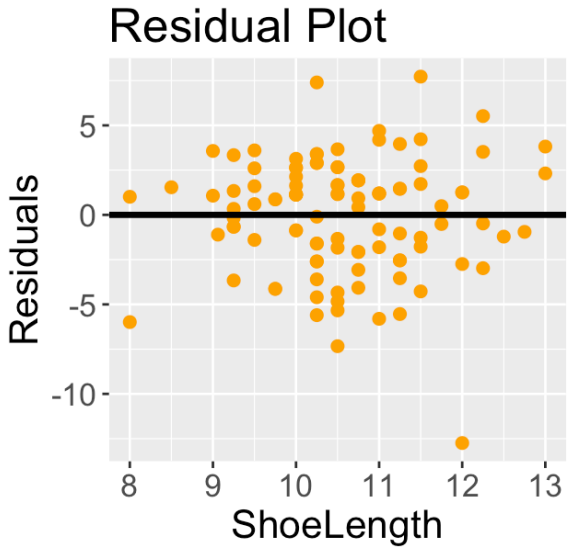
- Model assumptions
  - Check assumptions 1. SRS 2. Linearity 3. Constant SD 4. Normality
- Prediction
  - Mean response $\hat{\mu}_y = b_0 + b_1 x$
  - Individual response $\hat{y} = b_0 + b_1 x$
- Inference for predictions
  - Confidence interval for mean response $\hat{\mu}_y \pm t^* \text{SE}_{\hat{\mu}_y}$
  - Prediction interval for individual response $\hat{y} \pm t^* \text{SE}_{\hat{y}}$

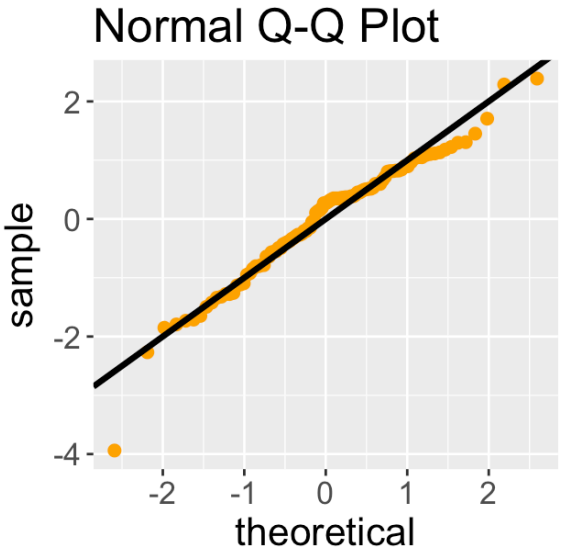
Check assumptions

1. The sample is an **SRS** from the population.
   - Check data collecting process.
2. There is a **linear** relationship between x and y.
   - Check scatterplot (linear) and residual plot (no pattern).
3. The **standard deviation** of the responses y about the population regression line is the **same** for all x.
   - Check residual plot: the spread of the residuals across the range of x should be roughly uniform.
4. The model residuals are **Normally** distributed.
   - Check Normal Q-Q plot: points should lie closely to the $y = x$ line.

Check assumptions: Height vs ShoeLength

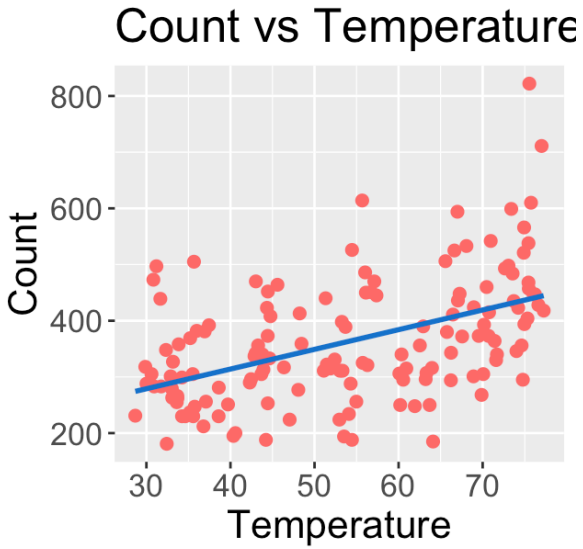


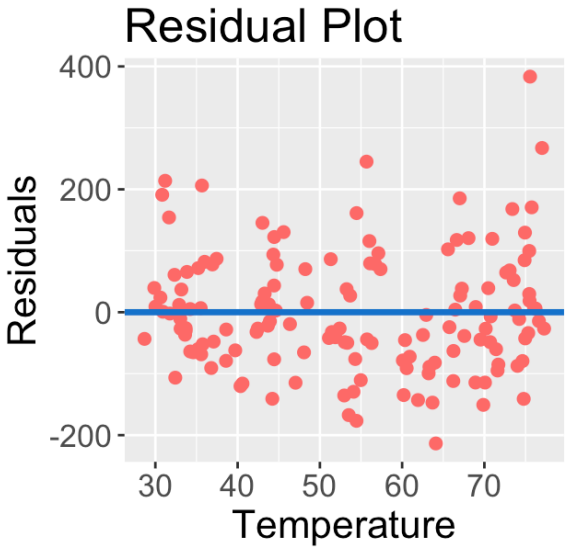


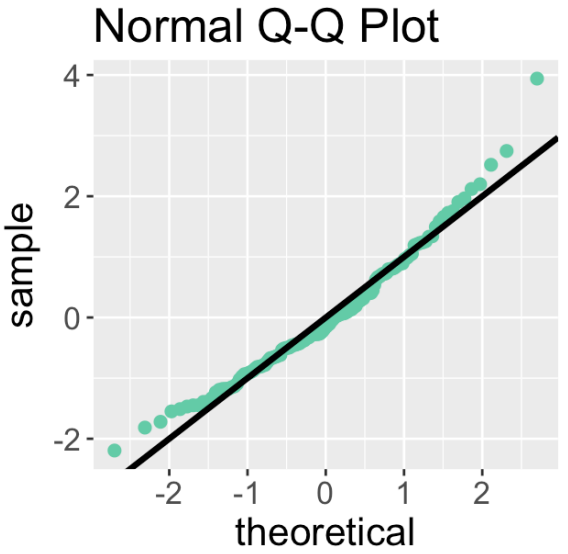


- Except for the two suspicious outliers, there is no clear violation of the linearity, constant SD and Normality assumptions.

Check assumptions: UFO vs Temperature







- Overall, the linearity, constant SD and Normality assumptions are approximately satisfied. However, there remains concern in the slightly curved scatterplot, residual plot and Normal Q-Q plot (possibly due to the outliers and other reasons).

When assumptions are violated...

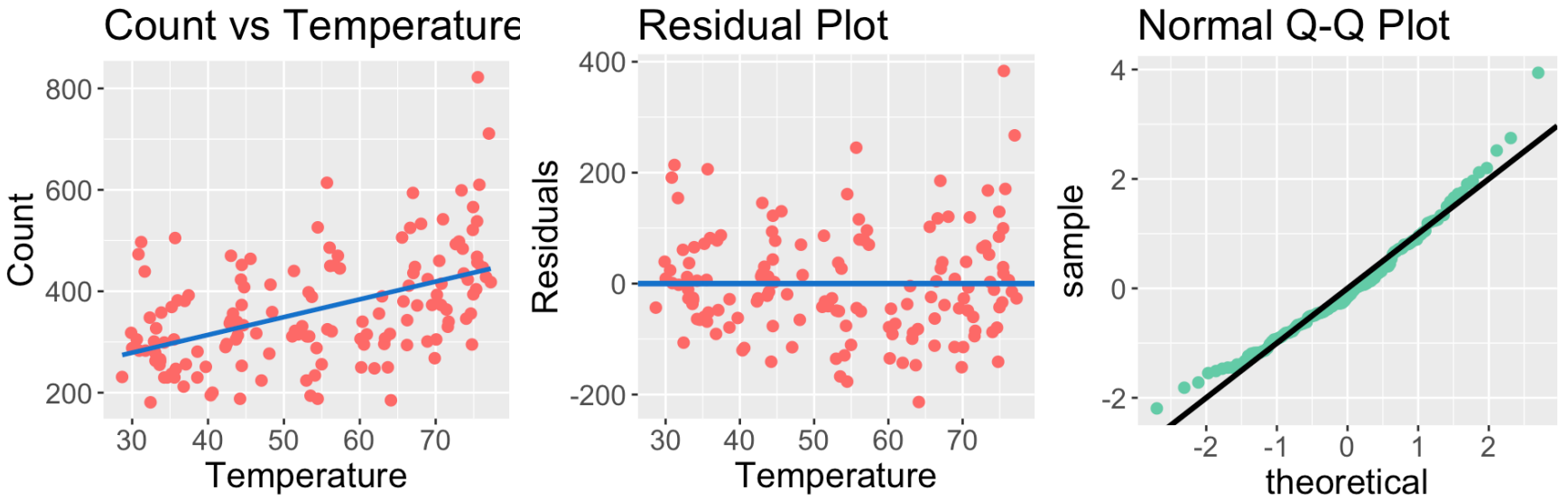
Violations in model assumptions usually lead to unreliable model inference. When assumptions are violated, we may try to

- Tranform the data.
- Remove outliers and unusual points.
- Use other statistical methods.

Note

- Try tranformation before removing outliers or unusual points. Because when data are tranformed, there could be no more outliers or unusual points.
- Be cautious to remove outliers and unusual points. Avoid data manipulation.
- Other statistical methods may include:
  - Polynomial regression (eg, $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$)
  - Regression assuming a non-Normal distribution (eg, Poisson regression)
  - Non-parametric regression (no assumption on distribution)

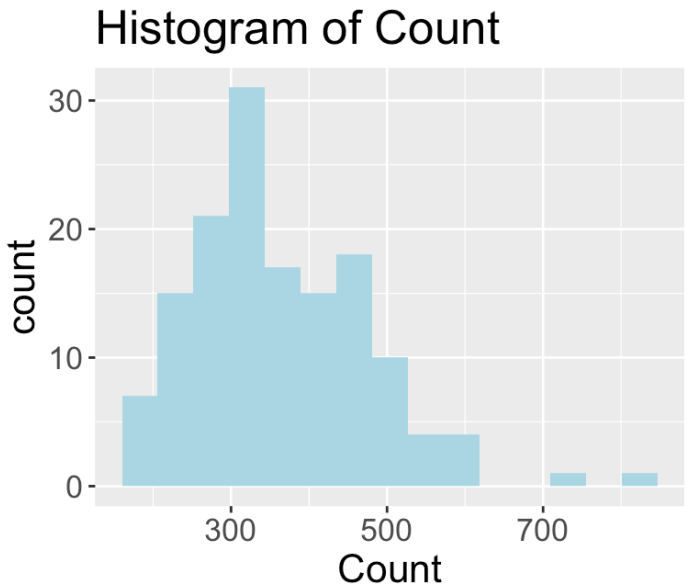
Transformation

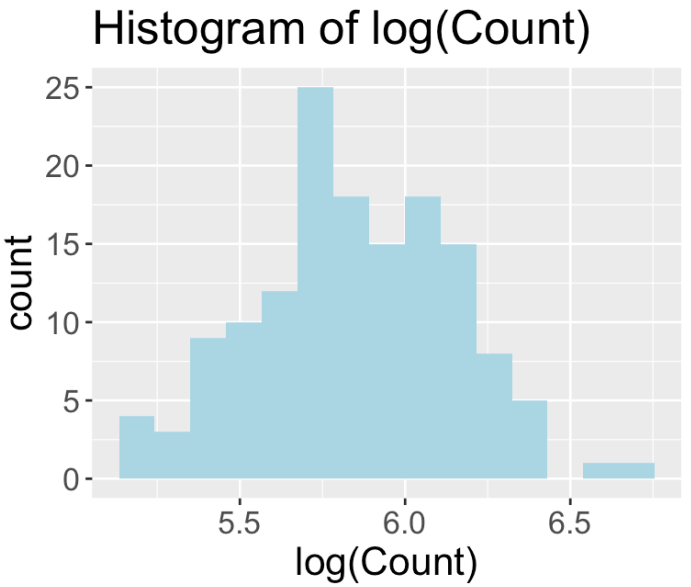


- Transformation in SLR usually applies a monotonic function (that preserves the order of the data) on the response variable Y or the explanatory variable X so that the model assumptions may be satisfied and model fitting may be improved.
- When data is of the type "count" (positive integers), transformation using the natrual logarithm function usually works very well.

Transformation - Distribution of *Count*

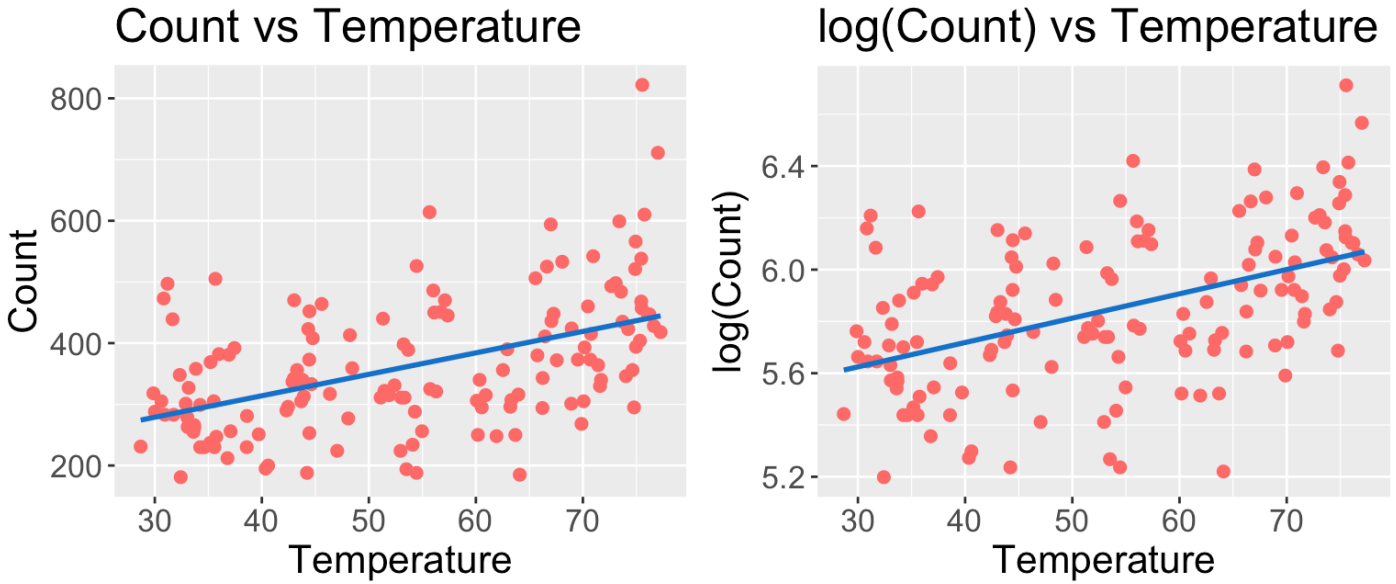
```
ggplot(UFO, aes(Count)) + geom_histogram(fill="lightblue", bins=15) +
  ggtitle("Histogram of Count")
ggplot(UFO, aes(log(Count))) + geom_histogram(fill="lightblue", bins=15) +
  ggtitle("Histogram of log(Count)")
```





- Before transformation, distribution of *Count* is right-skewed and has two extreme values. After transformation, distribution of *Count* is close to Normal.

Transformation - Scatterplot




```
cor(UFO$Temperature, UFO$Count); cor(UFO$Temperature, log(UFO$Count))
```

```
## [1] 0.4824087
```

```
## [1] 0.4814397
```

- After transformation, The two outliers become closer to the rest of the points.

Transformation - Model fitting and inferences

```
summary(m1 <- lm(Count ~ Temperature, data=UFO))
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 173.7714    29.9539   5.801 4.11e-08 ***
## Temperature   3.5055     0.5341   6.563 9.20e-10 ***
##
## Residual standard error: 97.64 on 142 degrees of freedom
## Multiple R-squared:  0.2327, Adjusted R-squared:  0.2273
## F-statistic: 43.07 on 1 and 142 DF,  p-value: 9.204e-10
```

- $\widehat{Count} = 173.8 + 3.5 \times Temp$

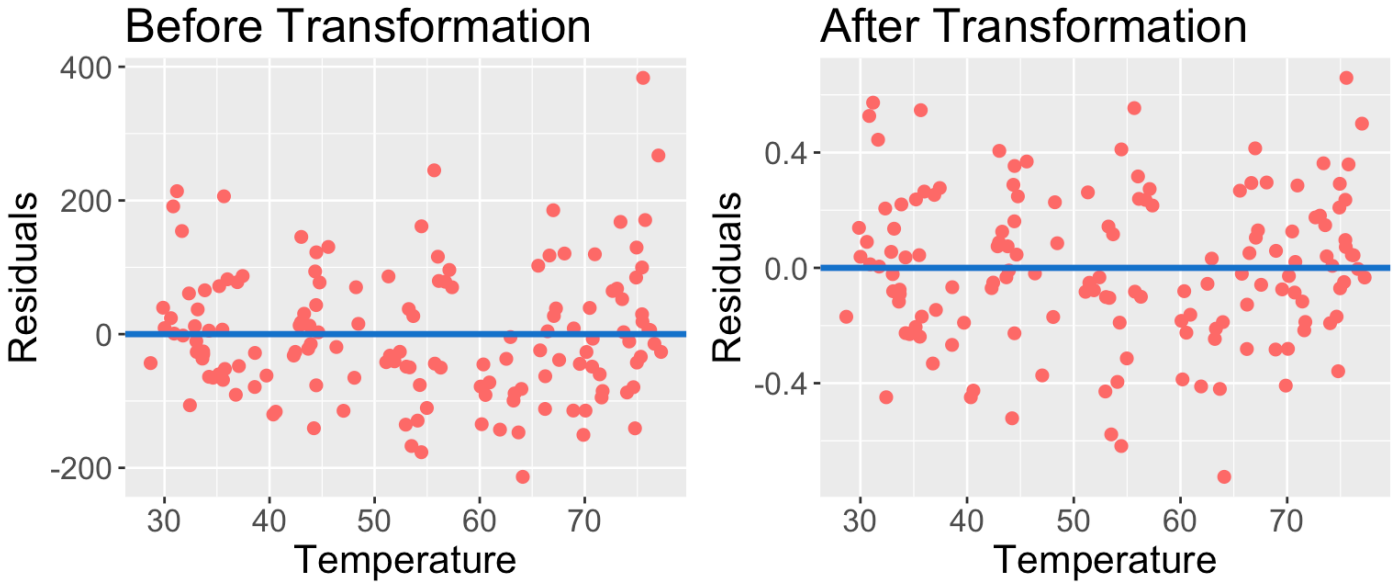
Transformation - Model fitting and inferences

```
summary(m2 <- lm(log(Count) ~ Temperature, data=UFO))
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.342374   0.080562  66.314  < 2e-16 ***
## Temperature 0.009403   0.001437   6.546 1.01e-09 ***
##
## Residual standard error: 0.2626 on 142 degrees of freedom
## Multiple R-squared:  0.2318, Adjusted R-squared:  0.2264
## F-statistic: 42.84 on 1 and 142 DF,  p-value: 1.005e-09
```

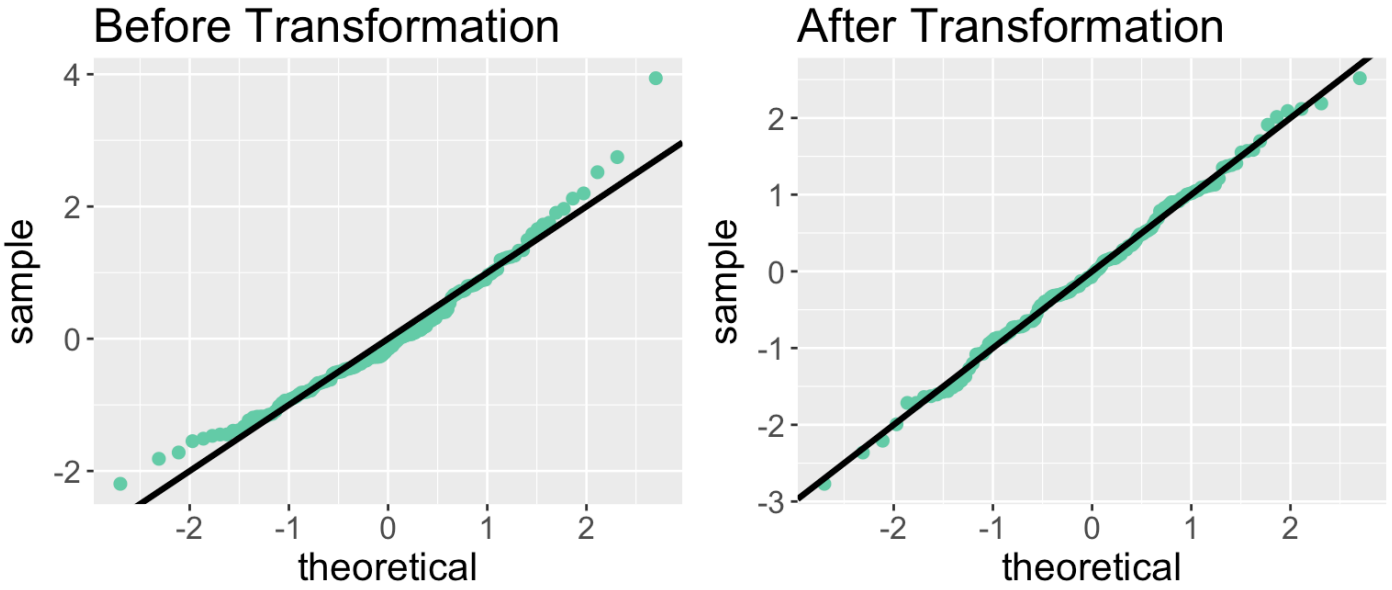
- $\widehat{\log(Count)} = 5.3 + 0.01 \times Temp$
- After transformation, we assume a linear relationship between *log(Count)* and *Temperature*. The significance and r^2 of the model stays similar as before.

Transformation - Residual plot



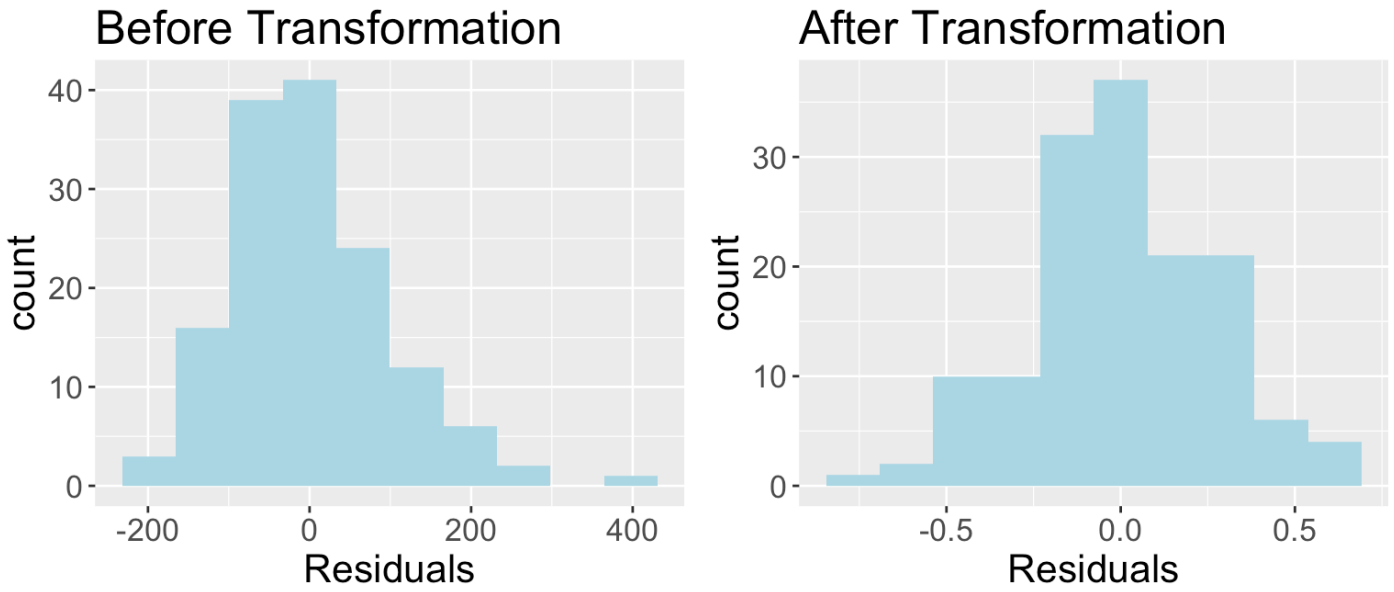
- After transformation, the pattern in the residual plot is less obvious and the spread of the residuals becomes more uniform for all the *Temperature* values. Also, the points are more evenly distributed above and below the $y = 0$ line.

Transformation - Normal Q-Q plot



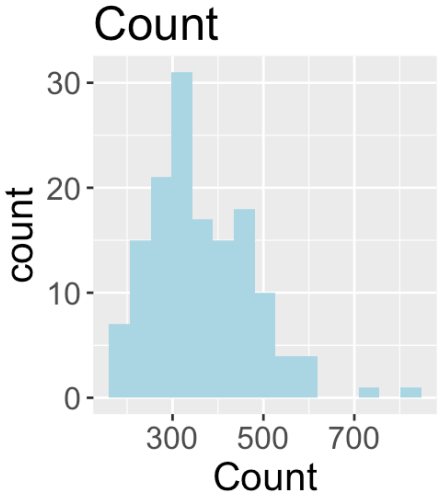
- After transformation, the points on the Normal Q-Q plot do not have curved trend anymore and form a quite straight line, and almost lie on the $y = x$ line, including the outliers.

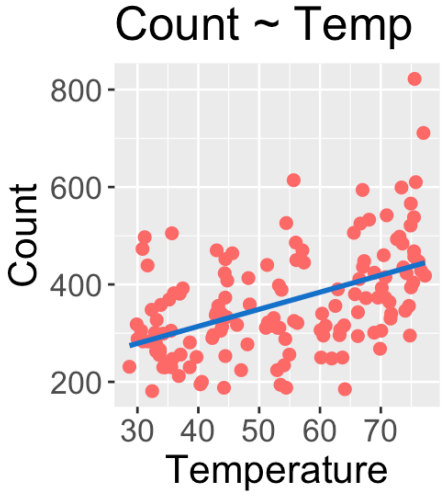
Transformation - Histogram of Residuals

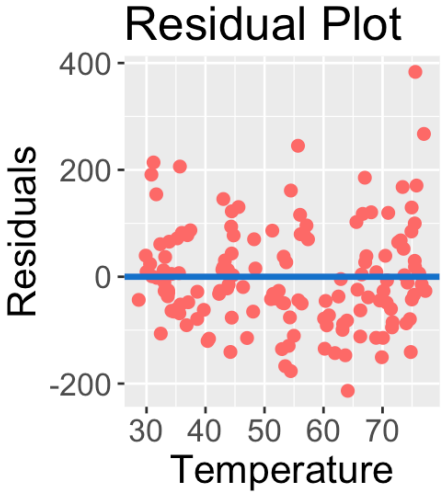


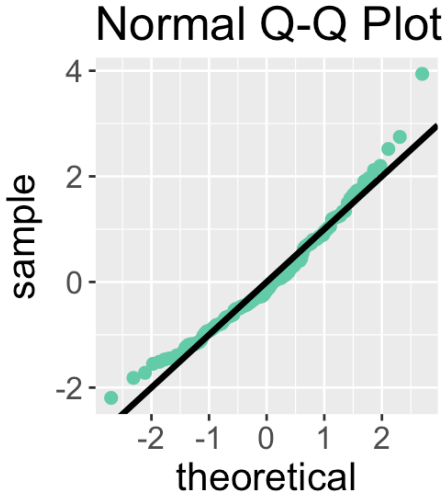
- Before transformation, the histogram of the residuals shows a right skewed distribution. After transformation, the distribution of residuals becomes almost symmetric and close to Normal.

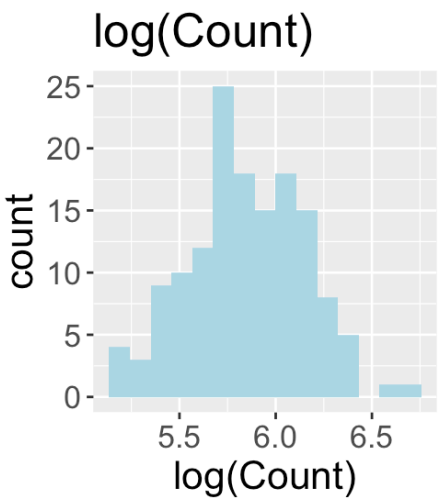
Transformation - Summary

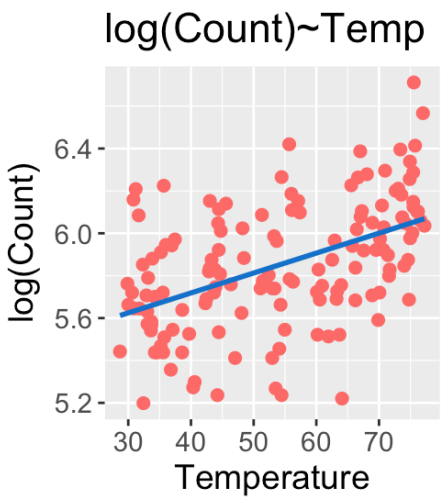


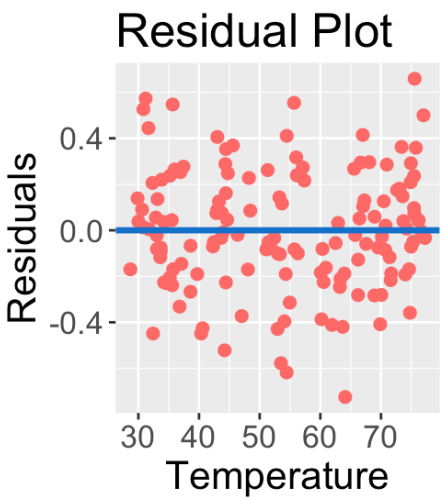


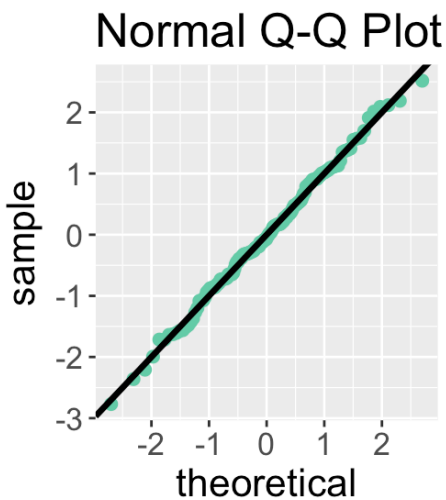












Transformation

- After transformation, all assumptions are satisfied now. The correlation coefficient, r^2 and significance of the model stay similar as before.
- What is the benefit of transformation?
- Let's do prediction

Prediction after transformation

```
predict(m1, list(Temperature=70), interval="confidence") # before transformation
```

```
##        fit     lwr      upr
## 1 419.1532 395.803 442.5034
```

- Before transformation $\hat{\mu}_{Count} = 419.2$ with 95% CI $[395.8, 442.5]$

```
predict(m2, list(Temperature=70), interval="confidence") # after transformation
```

```
##        fit      lwr      upr
## 1 6.000612 5.937811 6.063414
```

- After transformation $\hat{\mu}_{\log(Count)} = 6.0$ with 95% CI $[5.9, 6.1]$

```
exp(predict(m2, list(Temperature=70), interval="confidence")) # transform back
```

```
##        fit      lwr      upr
## 1 403.6759 379.1042 429.8403
```

- After transformation $\hat{\mu}_{Count} = e^{6.0} = 403.6$ with 95% CI $[e^{5.9}, e^{6.1}] = [379.1, 429.8]$

Prediction after transformation

```
predict(m1, list(Temperature=70), interval="prediction") # before transformation
```

```
##        fit      lwr      upr
## 1 419.1532 224.7331 613.5733
```

- Before transformation $\widehat{Count} = 419.2$ with 95% PI $[224.7, 613.6]$

```
predict(m2, list(Temperature=70), interval="prediction") # after transformation
```

```
##        fit      lwr      upr
## 1 6.000612 5.477712 6.523513
```

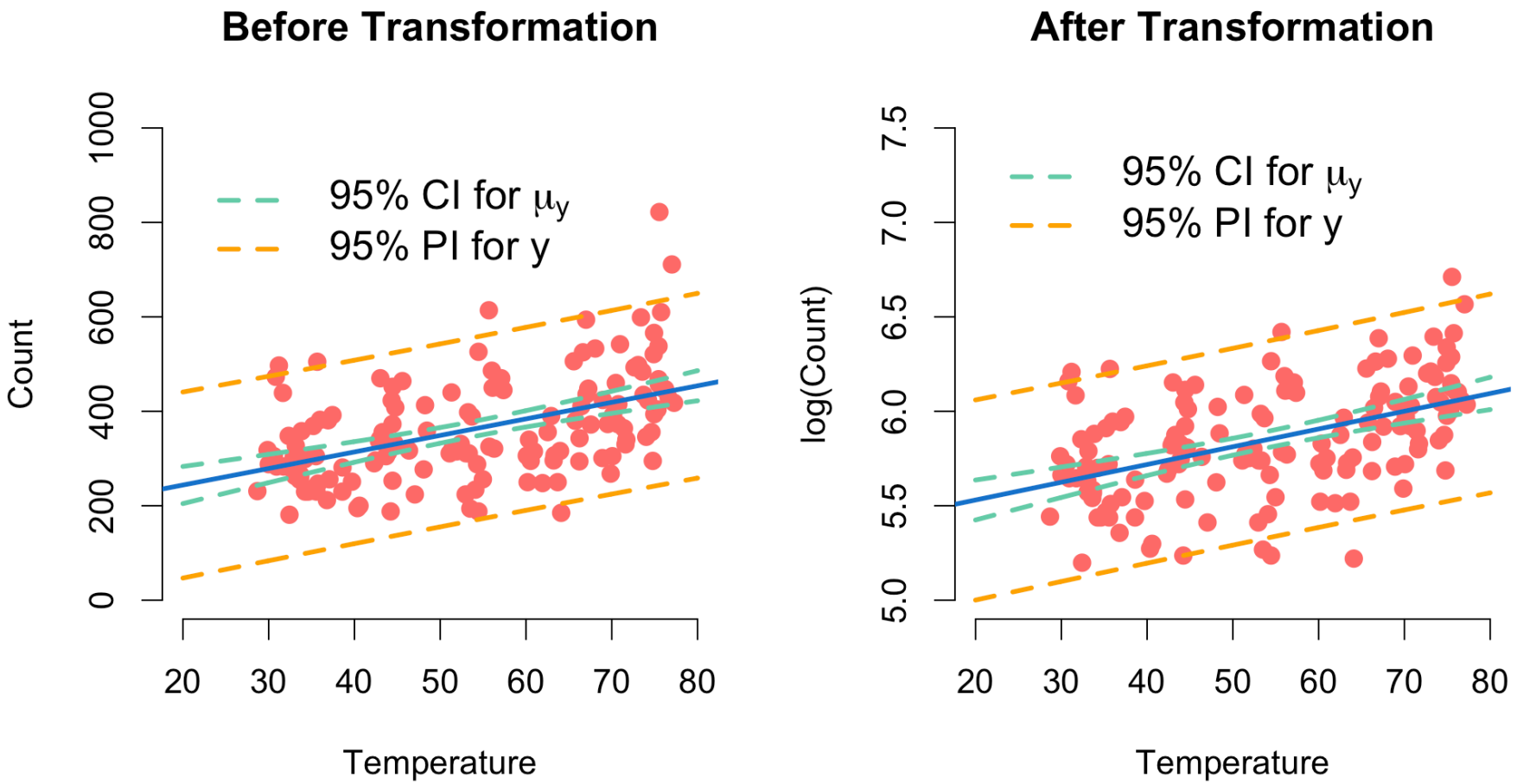
- After transformation $\widehat{\log(Count)} = 6.0$ with 95% PI $[5.5, 6.5]$

```
exp(predict(m2, list(Temperature=70), interval="prediction")) # transform back
```

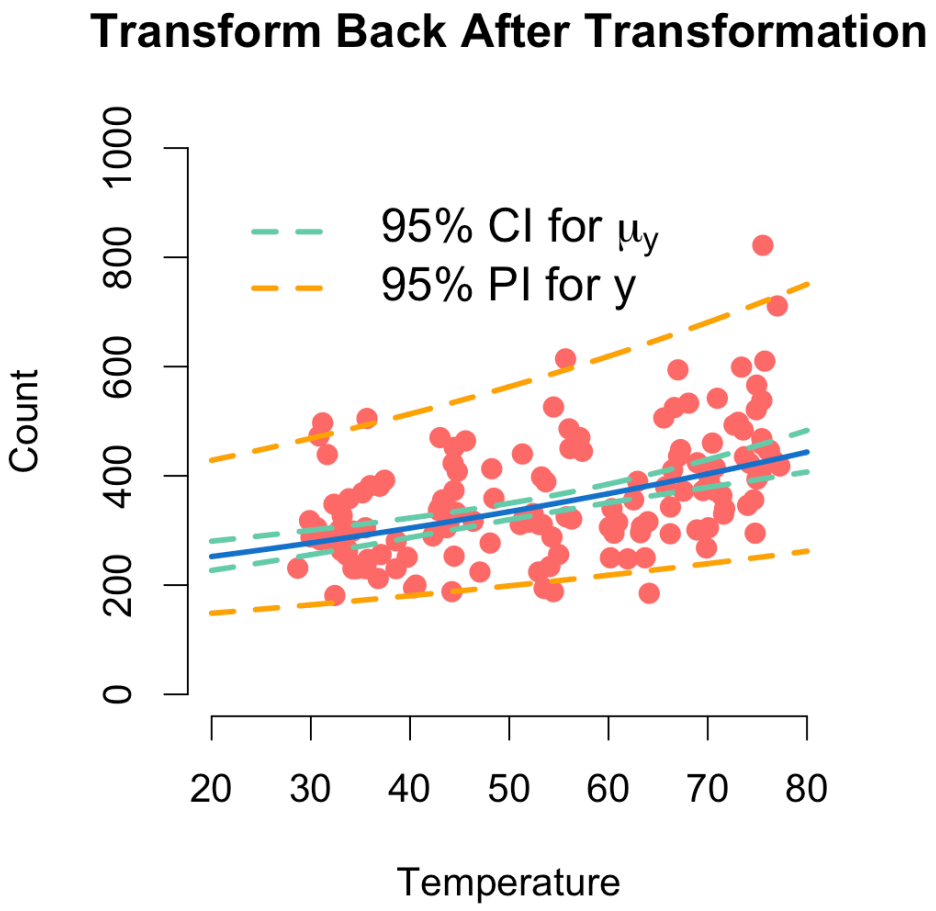
```
##        fit      lwr      upr
## 1 403.6759 239.2985 680.9665
```

- After transformation $\widehat{Count} = e^{6.0} = 403.6$ with 95% PI $[e^{5.5}, e^{6.5}] = [239.3, 680.1]$

Prediction after transformation



Prediction after transformation



- After transformation, all the data points lie within or close to the 95% PI lines suggesting that the prediction works better than before.
- If we transform the data back to the original scale after transformation, we find that the new model assumes a non-linear relationship between *Count* and *Temperature*, which fits the data better. The 95% PI lines are able to "capture" the non-linear pattern in the data, which again suggests better model fitting.

Notes

- The goal of transformation is to make the data satisify the model assumptions and/or to achieve better model fitting (more significant test, larger r^2 value, etc).
- Although this example in class does not improve r^2, after transformation, all assumptions are satisified and thus the new model explains the relationship between the two variables better than before.
- More importantly, when the model assumptions are satisfied, model inferences (for the intercept, slope and predictions) are much more reliable.
- Transformation in SLR can apply many other different functions on either the response variable or the explanatory variable. In Homework 11, you will transform the explanatory variable of a SLR model and then fit a polynomial regression model.

Summary

- Simple linear regression
  - Idea
  - Model $y = \mu_y + \epsilon = \beta_0 + \beta_1 x + \epsilon$ where $\epsilon \sim N(0, \sigma)$
- Inference for the regression line
  - Confidence intervals $b_0 \pm t^* \text{SE}_{b_0}$ and $b_1 \pm t^* \text{SE}_{b_1}$
  - Significance test $t = \frac{b_1 - 0}{\text{SE}_{b_1}} \overset{approx.}{\sim} t(n-2)$
- Model assumptions
  - Check assumptions 1. SRS 2. Linearity 3. Constant SD 4. Normaility
  - Tranformation and transforming back
- Prediction
  - Mean response $\hat{\mu}_y = b_0 + b_1 x$ with confidence interval $\hat{\mu}_y \pm t^* \text{SE}_{\hat{\mu}_y}$
  - Individual response $\hat{y} = b_0 + b_1 x$ with prediction interval $\hat{y} \pm t^* \text{SE}_{\hat{y}}$