Midterm 1 STAT632

Brandon Keck (netID qh9701)

```
library(ggplot2)
library(dplyr)
```

```
library(readr)
ice_cream <- read_csv("~/Documents/EastBay/Spring2025/Stat632Regression/ice_cream.csv")
```

```
# head(ice_cream)
```

Question 1.

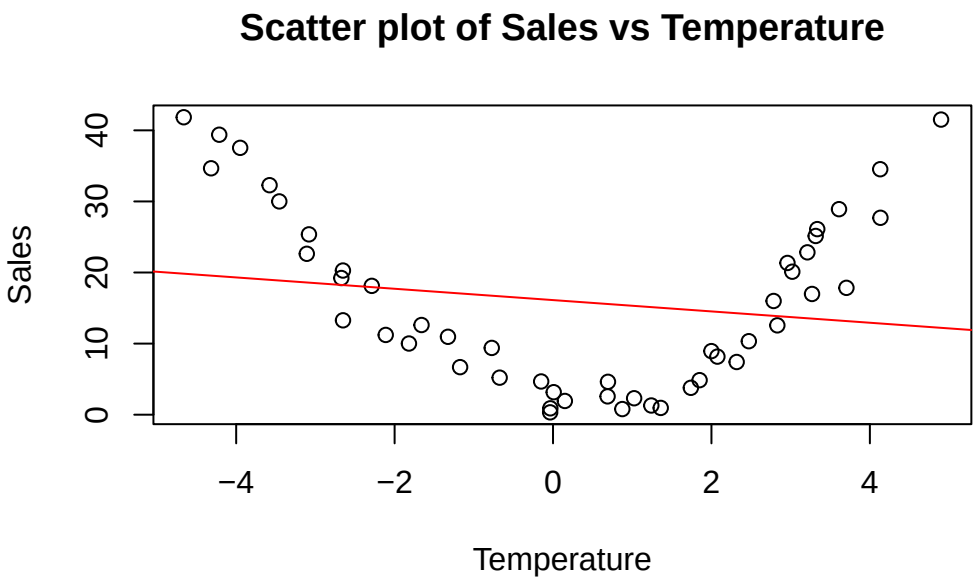
a)

Corrected

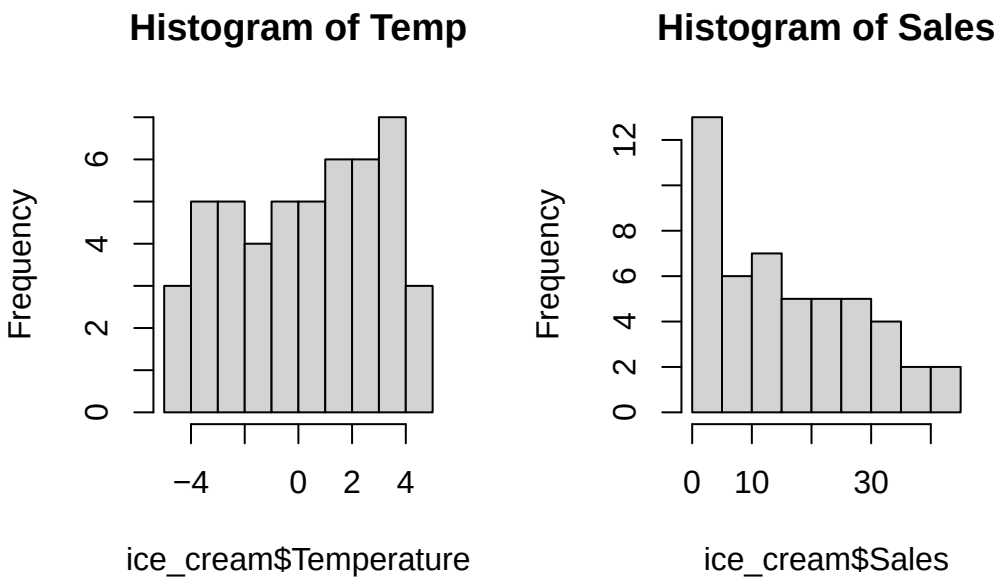
```
lm1 <- lm(Sales ~ Temperature, data = ice_cream)

plot(Sales ~ Temperature, data = ice_cream,
     xlab = "Temperature",
     ylab = "Sales",
     main = "Scatter plot of Sales vs Temperature")

abline(lm1, col = "red")
```




```
par(mfrow = c(1,2))
hist(ice_cream$Temperature, main = "Histogram of Temp")
hist(ice_cream$Sales, main = "Histogram of Sales")
```



Based on the scatter plot and the histograms the assumption of Linearity comes into question. We can see from the scatter plot that the data does not follow a linear pattern. In fact it looks like it might be a quadratic. The histogram of sales shows some right skew to it. The variance appears to be fairly constant.

Reflection:

I lost points here because I wasn't sure about how to do the Marginal Residuals plots. I thought I just had to do the Residuals vs Fitted plot which is a diagnostic tool but it doesn't show whether the residuals are skewed for specific predictor values.

b)

Corrected

```
lm2 <- lm(Sales ~ Temperature + I(Temperature^2), data = ice_cream)

summary(lm2)
```

```
Call:
lm(formula = Sales ~ Temperature + I(Temperature^2), data = ice_cream)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1543 -2.7190  0.2208  2.7906  5.5443

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.95177    0.70970   4.159 0.000138 ***
Temperature      -0.82468    0.17466  -4.722 2.22e-05 ***
I(Temperature^2)  1.82953    0.07403  24.715  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.264 on 46 degrees of freedom
Multiple R-squared:  0.9321,    Adjusted R-squared:  0.9292
F-statistic: 315.8 on 2 and 46 DF,  p-value: < 2.2e-16
```

```r
lm3 <- lm(Sales ~ poly(Temperature, 3), data=ice_cream)
lm4 <- lm(Sales ~ poly(Temperature, 4), data=ice_cream)
lm5 <- lm(Sales ~ poly(Temperature, 5), data=ice_cream)
lm6 <- lm(Sales ~ poly(Temperature, 6), data=ice_cream)
```

```r
summary(lm3)
```

```
Call:
lm(formula = Sales ~ poly(Temperature, 3), data = ice_cream)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3745 -2.4962 -0.1326  2.3011  5.7732

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            15.9053     0.4553  34.936  < 2e-16 ***
poly(Temperature, 3)1 -14.8858     3.1869  -4.671 2.73e-05 ***
poly(Temperature, 3)2  80.6755     3.1869  25.315  < 2e-16 ***
poly(Temperature, 3)3   5.7552     3.1869   1.806   0.0776 .
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.187 on 45 degrees of freedom
Multiple R-squared:  0.9367,    Adjusted R-squared:  0.9325
F-statistic:   222 on 3 and 45 DF,  p-value: < 2.2e-16
```

```r
summary(lm4)
```

```
Call:
lm(formula = Sales ~ poly(Temperature, 4), data = ice_cream)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0229 -2.0413  0.1652  2.1342  4.8582

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            15.9053     0.4337  36.677  < 2e-16 ***
poly(Temperature, 4)1 -14.8858     3.0356  -4.904 1.32e-05 ***
poly(Temperature, 4)2  80.6755     3.0356  26.577  < 2e-16 ***
poly(Temperature, 4)3   5.7552     3.0356   1.896   0.0645 .
poly(Temperature, 4)4  -7.1821     3.0356  -2.366   0.0225 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.036 on 44 degrees of freedom
Multiple R-squared:  0.9438,    Adjusted R-squared:  0.9387
F-statistic: 184.9 on 4 and 44 DF,  p-value: < 2.2e-16
```

```r
summary(lm5)
```

```
Call:
lm(formula = Sales ~ poly(Temperature, 5), data = ice_cream)

Residuals:
    Min      1Q  Median      3Q     Max
-8.5704 -1.8596  0.3919  2.2783  4.4607
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             15.905      0.433  36.731  < 2e-16 ***
poly(Temperature, 5)1  -14.886      3.031  -4.911 1.35e-05 ***
poly(Temperature, 5)2   80.675      3.031  26.616  < 2e-16 ***
poly(Temperature, 5)3    5.755      3.031   1.899   0.0643 .
poly(Temperature, 5)4   -7.182      3.031  -2.369   0.0224 *
poly(Temperature, 5)5   -3.222      3.031  -1.063   0.2937
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.031 on 43 degrees of freedom
Multiple R-squared:  0.9453,    Adjusted R-squared:  0.9389
F-statistic: 148.6 on 5 and 43 DF,  p-value: < 2.2e-16
```

```r
summary(lm6)
```

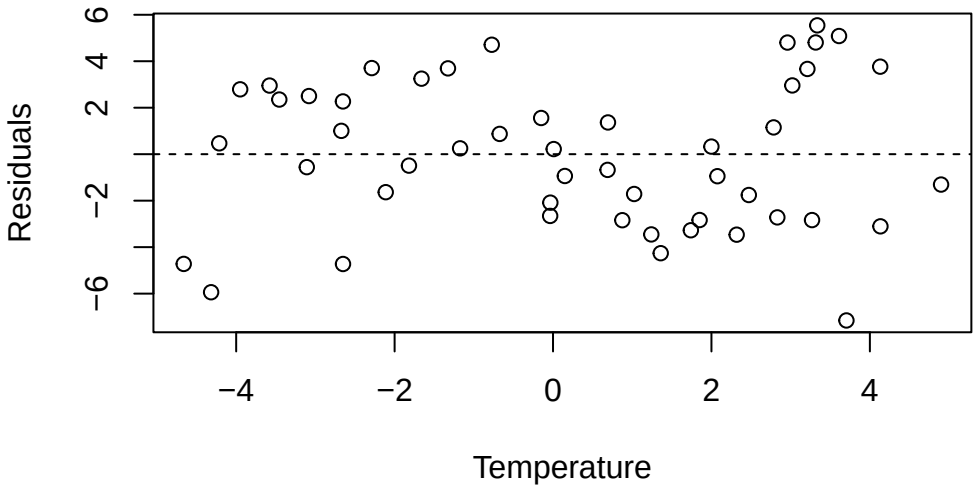
```
Call:
lm(formula = Sales ~ poly(Temperature, 6), data = ice_cream)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0327 -1.7822  0.4806  1.9979  4.5498

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            15.9053     0.4337  36.677  < 2e-16 ***
poly(Temperature, 6)1 -14.8858     3.0356  -4.904 1.45e-05 ***
poly(Temperature, 6)2  80.6755     3.0356  26.576  < 2e-16 ***
poly(Temperature, 6)3   5.7552     3.0356   1.896   0.0649 .
poly(Temperature, 6)4  -7.1821     3.0356  -2.366   0.0227 *
poly(Temperature, 6)5  -3.2221     3.0356  -1.061   0.2946
poly(Temperature, 6)6  -2.8352     3.0356  -0.934   0.3557
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.036 on 42 degrees of freedom
Multiple R-squared:  0.9464,    Adjusted R-squared:  0.9387
F-statistic: 123.6 on 6 and 42 DF,  p-value: < 2.2e-16
```

```r
plot(ice_cream$Temperature, resid(lm2),
     xlab = "Temperature",
     ylab = "Residuals")
abline(h = 0, lty = 2)
```



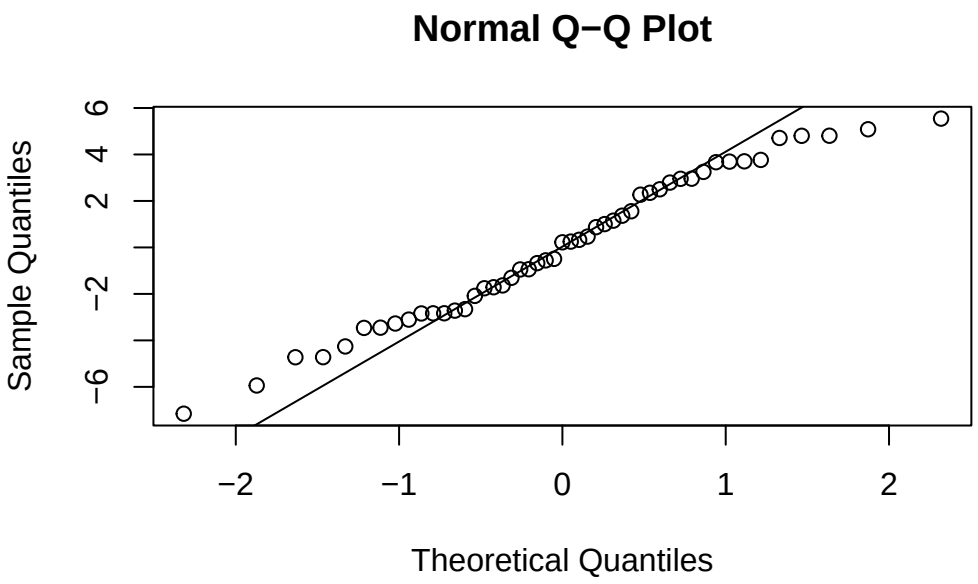
Since the scatterplot suggested a non-linear relationship, I tested six polynomial models as we had previously done in class to determine the best fit. After comparing the Adjusted R-squared values the second order polynomial model (lm2) significantly improved model performance. The residual plot confirms that a second-degree polynomial sufficiently captures the data pattern.

Reflection:

I originally tested many polynomial models but did not explain why I chose the final model (lm2). I also didn't explicitly compare the adjusted R-squared values or residuals plots.

c)

```
qqnorm(resid(lm2))
qqline(resid(lm2))
```



From the Normal QQ plot we see that most of the data follows the line. There is some curvature at both ends of the tails however this looks like we are on the right track in transforming our data. Therefore the assumption of Normality appears to be satisfied.

d)

Simple linear regression model for the population

$$Y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon_i$$

e)

```
ind <- which(abs(rstandard(lm2)) > 2)

ice_cream[ind, ]
```

```
# A tibble: 1 x 2
  Temperature Sales
        <dbl> <dbl>
1        3.70  17.8
```

It seems that there is only two outliers within our new Polynomial model. However, we have no indication that these are incorrect and so we have no motivation to remove them.

Question 2.

a)

```
summary(lm2)
```

```
Call:
lm(formula = Sales ~ Temperature + I(Temperature^2), data = ice_cream)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1543 -2.7190  0.2208  2.7906  5.5443

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.95177    0.70970   4.159 0.000138 ***
Temperature      -0.82468    0.17466  -4.722 2.22e-05 ***
I(Temperature^2)  1.82953    0.07403  24.715  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.264 on 46 degrees of freedom
Multiple R-squared:  0.9321,    Adjusted R-squared:  0.9292
F-statistic: 315.8 on 2 and 46 DF,  p-value: < 2.2e-16
```

From the R output we can see that all variables included in our model are useful in explaining Sales our variable of interest. We have an R^2 of 93% mean that 93% of the variability of Sales can be explained by Temperature

b)

Estimated Regression Equation:

$$\hat{y_i} = \hat{\beta_0} - \hat{\beta_1} x_1 + \hat{\beta} x_2^2$$

$$\boxed{\hat{Sales} = 2.95177 - 0.82468(Temperature) + 1.82953(Temperature^2)}$$

c)

The intercept represents the predicted Sales when Temperature is 0. If the temperature were to be 0 the model predicts Sales to be \$2.95. Typically when we interpret the intercept there is no value in the context of the problem because our predictor variables aren't typically 0. For instance the Fandango movie rating it wouldn't be logical to consider an IMDb rating of 0 because ratings are typically from 1-5. However, ice cream can be 0 or below. Reviewing the data itself there are negative values and some values that are even essentially 0. So in the context of this problem the interpretation is meaningful.

d)

$$\hat{Sales} = 2.95177 - 0.82468(2) + 1.82953(2^2)$$

```
2.95177 -0.82468 * (2) + 1.82953 * (2^2)
```

```
[1] 8.62053
```

From the formula we calculated the point estimate for the sales unit of ice cream when the temperature is 2 degrees Celsius is approximately 8.621

```
pred <- data.frame(Temperature = 2)
predict(lm2, newdata = pred, interval = "prediction")
```

```
       fit      lwr    upr
1 8.620516 1.938028 15.303
```

We are 95% confident that the Sales units of ice cream when the temperature is 2 degrees Celsius will be 8.621 and that the sales units will fall between 1.94 and 15.303.

e)

Corrected

```r
new_x <- data.frame(Temperature = 2)
predict(lm2, newdata = new_x, interval = "prediction", level = 0.97)
```

```
       fit      lwr      upr
1 8.620516 1.185756 16.05528
```

So our predicted value is still 8.621 when the temperature is 2 degrees C. However, we now see that our interval has widened even more to 1.186 to 16.055.

For a temperature of 2 degrees C, the predicted sales are 8.62 units. Additionally, the 97% prediction interval is (1.19, 16.06). This means that if the temperature is 2 degrees C, the actual number of ice cream sales is likely to be between 1.19 and 16.06 units.

Reflection:

I got this problem wrong because I made the wrong interpretation of the output. I had accidentally used the wrong flashcard, which contained the incorrect response. This led me to misinterpret the prediction interval.