Lab 10: Quadratic Regression

Electricity Demand

We have data from the Australian Energy Market Operator and the Australian Bureau of Meteorology with daily electricity demand for Victoria, Australia, in 2014. For each day, we have:

- `Demand`: Total electricity demand in GW for Victoria, Australia
- `WorkDay`: "WorkDay" for work days, and "Other" for non work days
- `Temperature`: The daily high temeprature in degrees Celsius

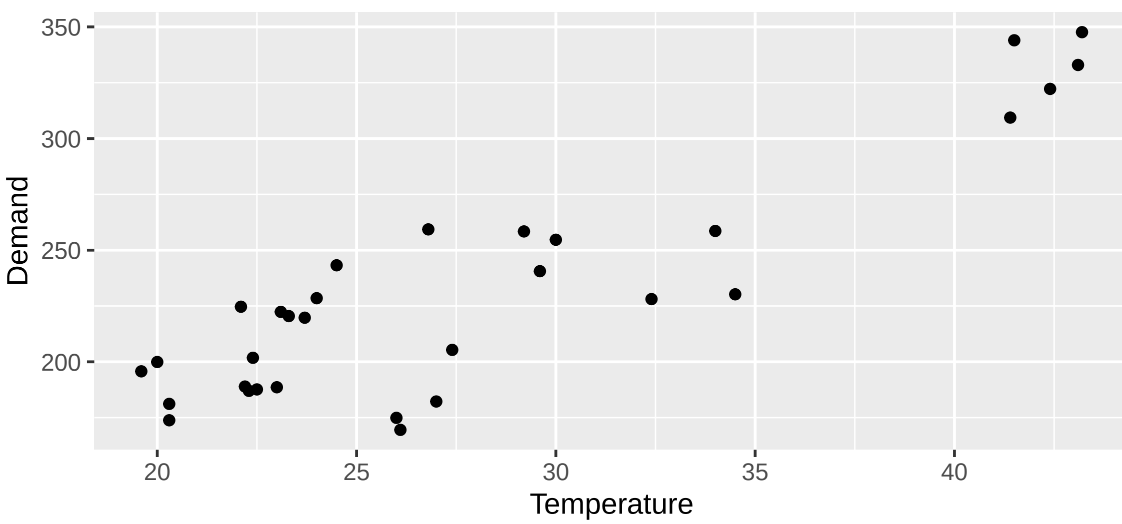
```
## # A tibble: 6 x 3
##   Demand WorkDay Temperature
##    <dbl> <chr>         <dbl>
## 1   175. Other          26
## 2   189. WorkDay        23
## 3   189. WorkDay        22.2
## 4   174. Other          20.3
## 5   170. Other          26.1
## 6   196. WorkDay        19.6
```

As always with data collected over time, we should be suspicious of the condition of independence. For today, let's set that aside and focus on an analysis of the relationships between these variables.

The `elecdaily_jan` data frame contains the data for just January, and the `elecdaily` data frame contains the data for the full year.

1. Make a plot of the data for January (`elecdaily_jan`), treating Demand as the response and Temperature as the explanatory variable.

```
ggplot(data = elecdaily_jan, mapping = aes(x = Temperature, y = Demand)) +
  geom_point()
```



2. Fit a linear regression model using Temperature as an explanatory variable and Demand as the response. Print a summary of your model fit.

```
lm_fit <- lm(Demand ~ Temperature, data = elecdaily_jan)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Demand ~ Temperature, data = elecdaily_jan)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -50.470 -10.333   1.915  18.520  35.012
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  59.3294    17.2626   3.437   0.0018 **
## Temperature   6.1554     0.5963  10.322  3.2e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.43 on 29 degrees of freedom
## Multiple R-squared:  0.786,  Adjusted R-squared:  0.7787
## F-statistic: 106.5 on 1 and 29 DF,  p-value: 3.202e-11
```

3. Write down the equation for the estimated mean electricity demand as a function of temperature.

$\hat{mu} = 59.33 + 6.16 Temperature$

4. Find the predicted electricity demand from your model if the Temerature is 10 degrees C. Do you trust your prediction?

```r
59.33 + 6.16 * 10
```

```
## [1] 120.93
```

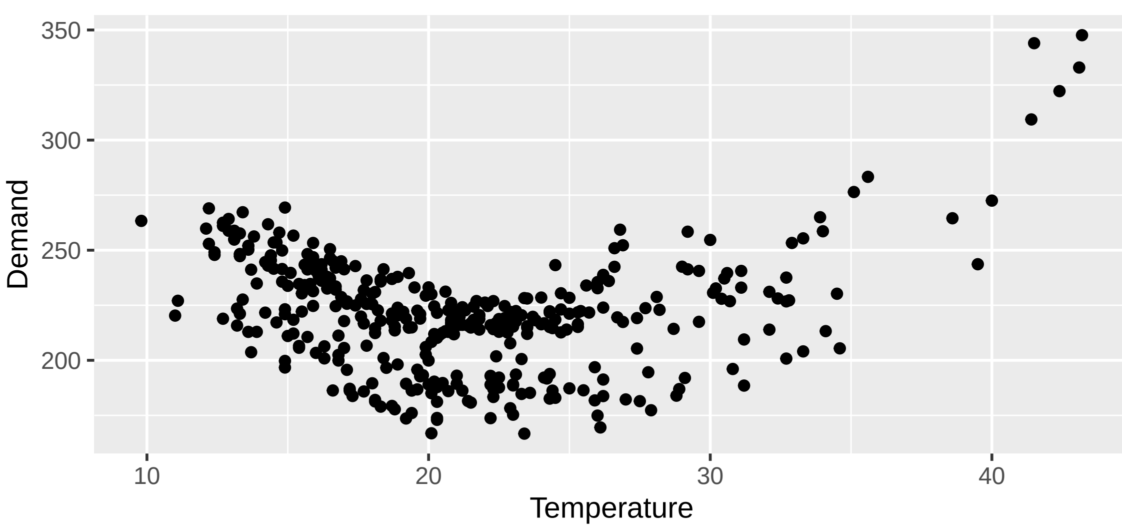
```r
predict(lm_fit, newdata = data.frame(Temperature = 10))
```

```
##        1
## 120.8832
```

We should not trust this prediction because the `elecdaily_jan` data set has only observations with temperatures between about 20 degrees and 43 degrees. It's not reliable to extrapolate a linear relationship beyond the observed range of the data.

5. Create a plot of the data for the full year, in the `elecdaily` data frame. How did your prediction from part 4 do?

```r
ggplot(data = elecdaily, mapping = aes(x = Temperature, y = Demand)) +
  geom_point()
```



Not very well! There was actually a quadratic relationship between temperature and demand, and that was not captured in the model from part 2 that we used to get the prediction.

6. Fit a quadratic regression model to the data for the full year and print out the model summary.

```
quad_fit <- lm(Demand ~ poly(Temperature, degree = 2, raw = TRUE), data = elecdaily)
summary(quad_fit)
```

```
##
## Call:
## lm(formula = Demand ~ poly(Temperature, degree = 2, raw = TRUE),
##     data = elecdaily)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -45.42 -18.03   6.11  14.43  48.72
##
## Coefficients:
##                                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                387.69194   10.73907   36.10   <2e-16 ***
## poly(Temperature, degree = 2, raw = TRUE)1 -15.28379    0.91830  -16.64   <2e-16 ***
## poly(Temperature, degree = 2, raw = TRUE)2   0.32371    0.01861   17.39   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.62 on 362 degrees of freedom
## Multiple R-squared:  0.4602, Adjusted R-squared:  0.4573
## F-statistic: 154.3 on 2 and 362 DF,  p-value: < 2.2e-16
```

7. Write down the equation for the estimated mean electricity demand as a function of temperature.

$\hat{\mu} = 387.70 - 15.28 Temperature + 0.32 Temperature^2$

8. Find the predicted electricity demand from your model if the Temperature is 10 degrees C.

```
387.70 - 15.28 * 10 + 0.32 * 10^2
```

```
## [1] 266.9
```

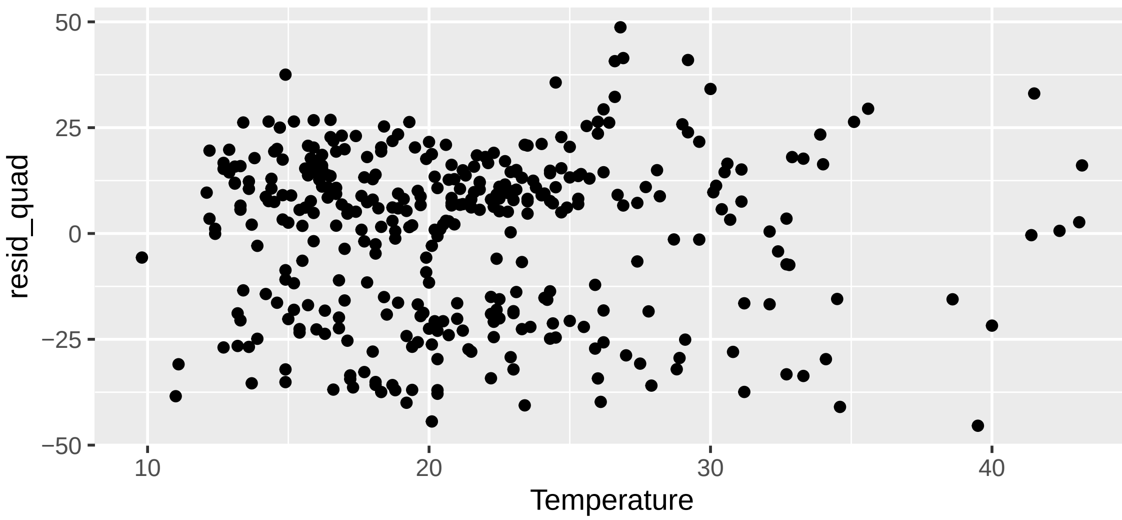
```
predict(quad_fit, newdata = data.frame(Temperature = 10))
```

```
##        1
## 267.2251
```

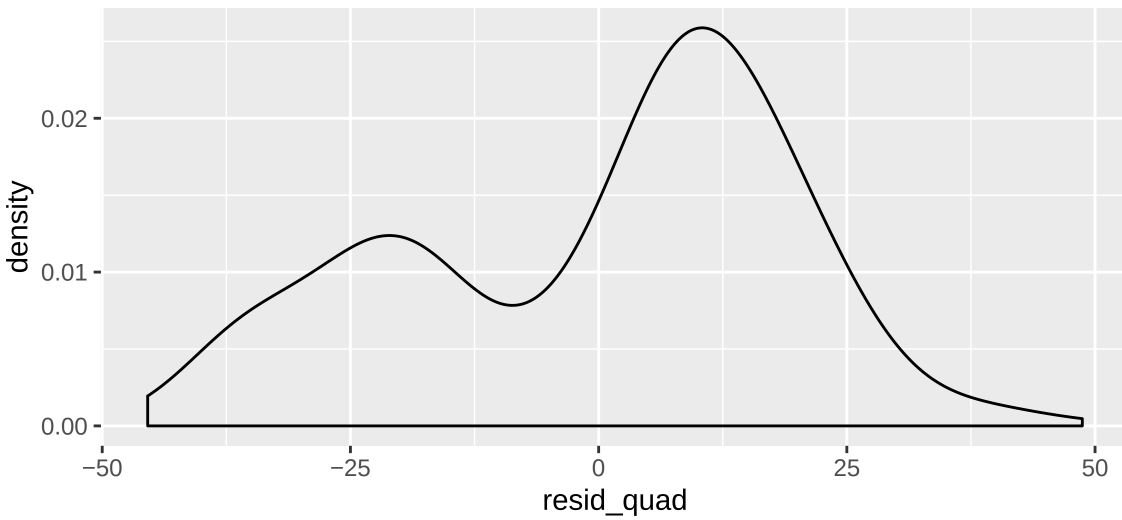
9. Make some residual diagnostic plots from your quadratic regression model. Do you see any evidence of problems?

```
elecdaily <- elecdaily %>%
  mutate(
    resid_quad = residuals(quad_fit)
  )

ggplot(data = elecdaily, mapping = aes(x = Temperature, y = resid_quad)) +
  geom_point()
```



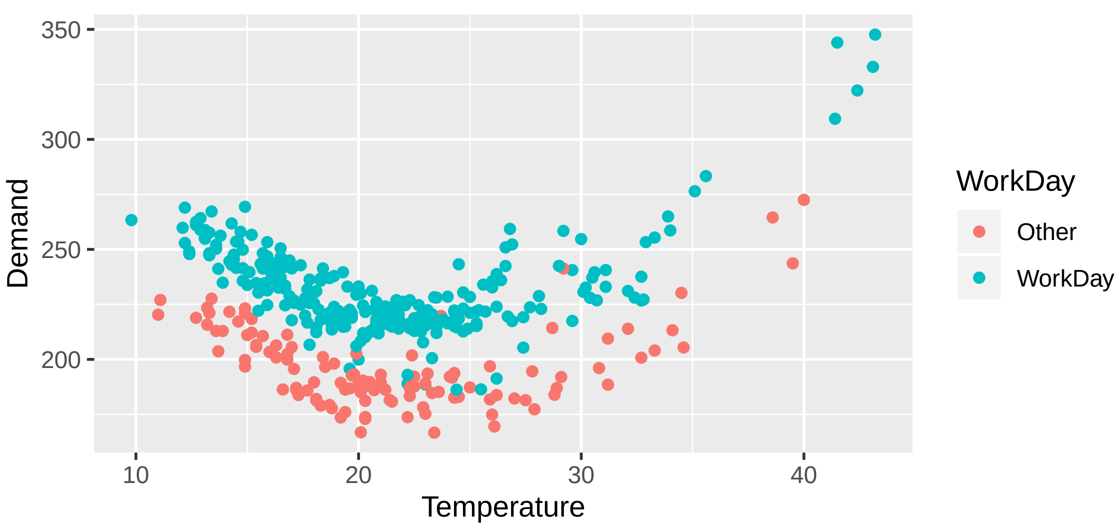

```r
ggplot(data = elecdaily, mapping = aes(x = resid_quad)) +
  geom_density()
```



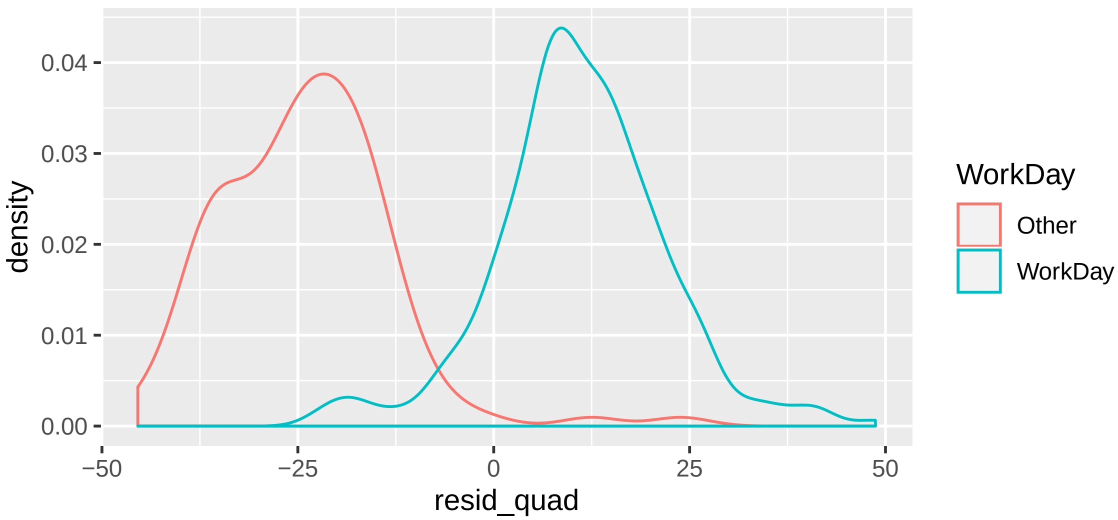
The residual density plot is bimodal. This suggests we should investigate further as there may be two groups in our data set.

10. Make another plot of the data, this time coloring each day according to whether it is a work day or not. What's going on?

```r
ggplot(data = elecdaily, mapping = aes(x = Temperature, y = Demand, color = WorkDay)) +
  geom_point()
```



```r
ggplot(data = elecdaily, mapping = aes(x = resid_quad, color = WorkDay)) +
  geom_density()
```



The electricity demand on work days is consistently higher than the electricity demand on non work days that have the same temperature.