Report 2

Lab2 GroupA

2025-02-07

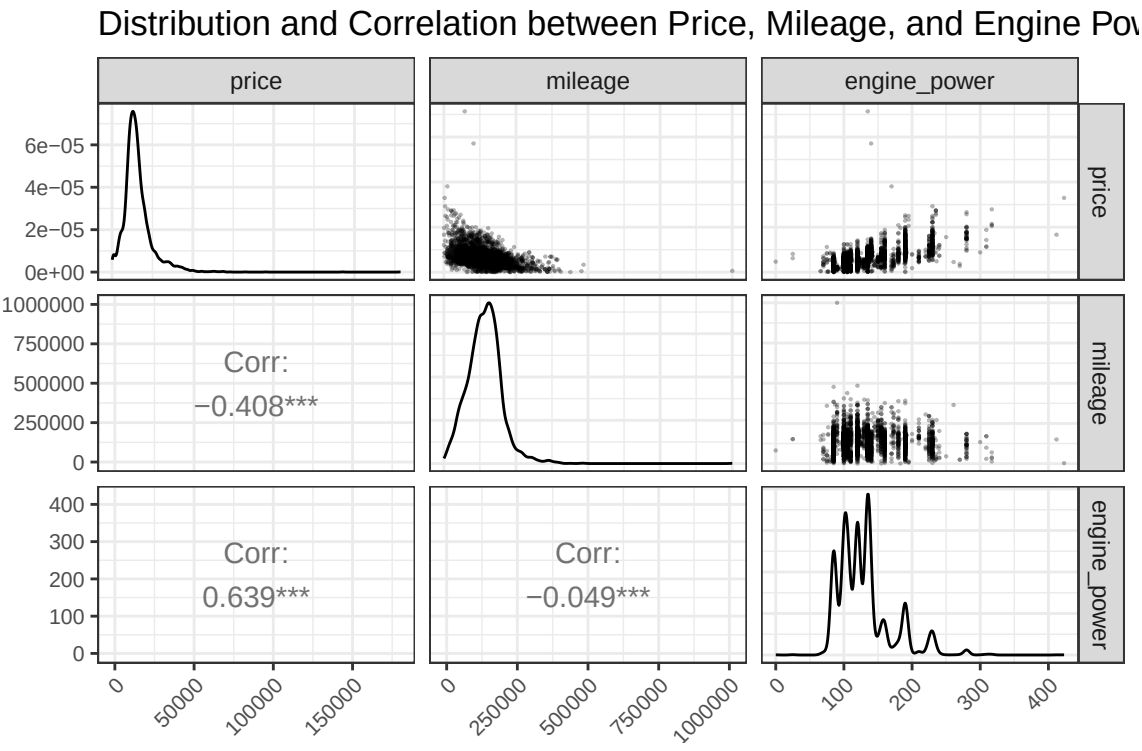
```
### load the library
library(dplyr)
library(ggplot2)
library(cowplot) #to arrange ggplot
library(GGally)
library(MASS) #for boxcox
### load the data
dat_bmw <- read.csv("BMWpricing_updated.csv")
```

address missing data

```
### Assigning NA to data points with negative mileage
dat_bmw[dat_bmw$mileage < 0, "mileage"] <- NA
### Converting mileage, engine power, and price into numeric
dat_bmw[,c(3,4,17)] <- apply(dat_bmw[,c(3,4,17)], 2, as.numeric)
### Create a data frame only has the variable of interests
dat_bmw2 <- dat_bmw[,c("price", "mileage", "engine_power")]
### Drop the data point with NA
dat_bmw2 <- dat_bmw2[complete.cases(dat_bmw2),]
# summary(dat_bmw)
```

scatterplot matrix

```
ggpairs(dat_bmw2,
        upper = list(continuous = wrap("points", alpha = 0.3, size = 0.1)),
        lower = list(continuous = wrap("cor", size = 4))) +
  labs(title = "Distribution and Correlation between Price, Mileage, and Engine Power") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 8),
        axis.text.y = element_text(size = 8))
```



scatter plot of price and mileage We first set the negative mileage (in one sample) as the missing data and excluded it in our following analysis.

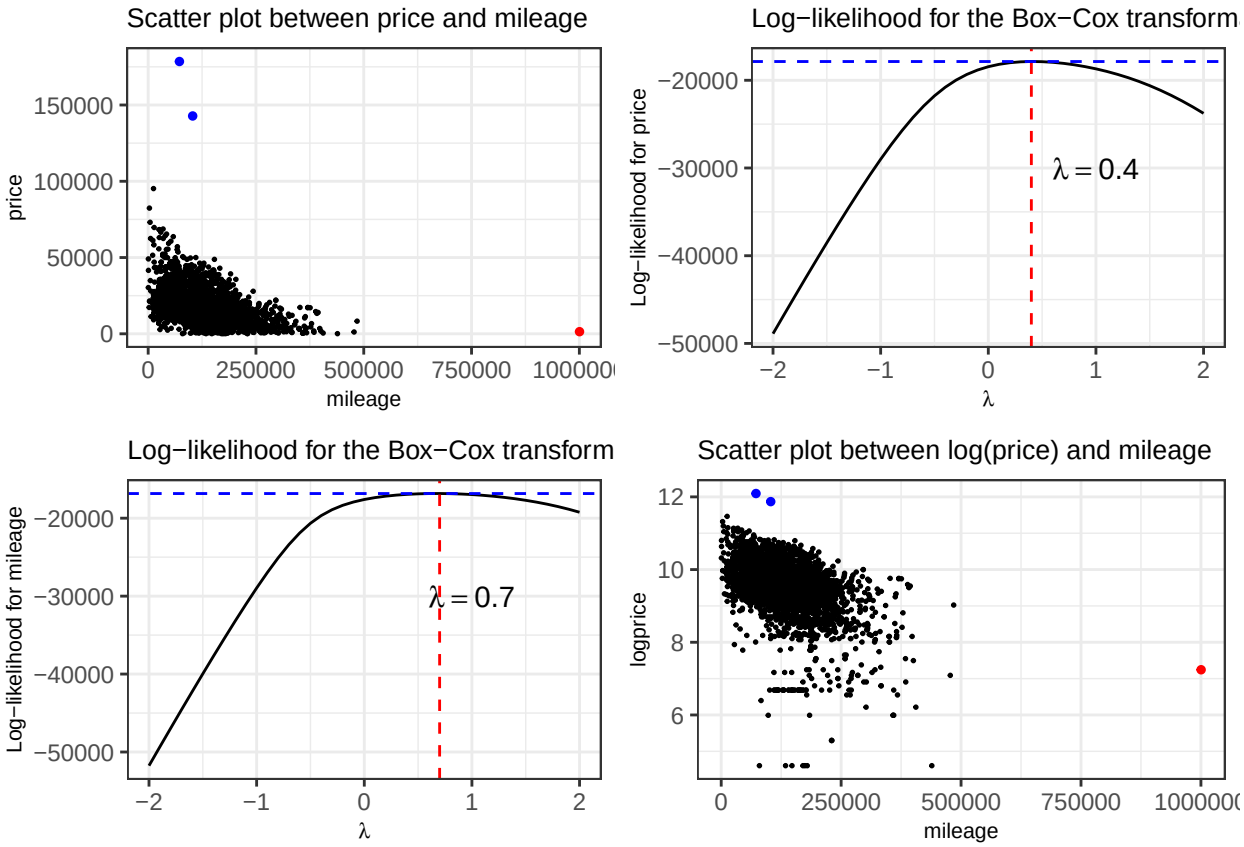
Our scatter plot shows a negative correlation between price and mileage, that is price tends to decrease with the longer mileage. However, this relationship is not strongly linear and a straight-line regression model may not be the best option. We find some problems in our data: 1) Both price and mileage follow highly right-skewed distributions; 2) There are some high leverage points on the right side of the scatter plot, especially the red point with around 1000000 miles; 3) Outliers with extreme high sold prices are observed on the top-left of the plot, especially these blue points with price over 100000.

We further transform price based on the log-likelihood for the Box-Cox transformation method. Though the λ with the greatest log-likelihood ($\lambda_{price} = 0.4, \lambda_{mileage} = 0.4$) and its 95% CI do not include 0, 0.5, and 1, there are slight differences in log-likelihood for price when $\lambda \in [0, 0.5]$ and for mileage when $\lambda \in [0.5, 1]$. For interpretability, we finally apply log-transformation on the price and keep the mileage in the data.

In the scatter plot of log(price) and mileage, we observe some potential outliers with low prices and short mileage.

```
# Scatter plot of price vs. mileage with highlighted outliers
scatterPlotPriceMileage <- ggplot(data = dat_bmw2, mapping = aes(x = mileage, y = price)) +
  geom_point(pch=19, cex=0.3) +  # Default points
  geom_point(data = subset(dat_bmw2, mileage > 500000), mapping = aes(x = mileage, y = price), pch=19,
  geom_point(data = subset(dat_bmw2, price > 100000), mapping = aes(x = mileage, y = price), pch=19, ce
  labs(title = "Scatter plot between price and mileage") +
  theme_bw() + theme(title = element_text(size=8))

# Box-Cox transformation for the price variable
boxcox_price <- boxcox(lm(dat_bmw2$price ~ 1), plotit = F)
boxcox_priceDF <- data.frame("lambda" = boxcox_price$x, "ll" = boxcox_price$y)
maxlabmda_price <- boxcox_price$x[which.max(boxcox_price$y)]

# Box-cox Plot for the price variable
boxcoxPlotPrice <- ggplot(data = boxcox_priceDF, mapping = aes(x = lambda, y = ll)) +
```

```r
  geom_line(color = "black", linewidth = 0.5) +
  geom_vline(xintercept = maxlabmda_price, linetype = "dashed", color = "red") +
  geom_hline(yintercept = max(boxcox_price$y) - qchisq(0.95, df = 1)/2, linetype = "dashed", color = "bl
  labs(title = "Log-likelihood for the Box-Cox transformation",
       x = expression(lambda),
       y = "Log-likelihood for price") +
  annotate("text", label = paste0("lambda ==", maxlabmda_price), x = 1, y = -30000, parse = T) +
  theme_bw() + theme(title = element_text(size=8))

# Box-cox transformation for the mile variable
boxcox_mile <- boxcox(lm(dat_bmw2$mileage ~ 1), plotit = F)
boxcox_mileDF <- data.frame("lambda" = boxcox_mile$x, "ll" = boxcox_mile$y)
maxlabmda_mile <- boxcox_mile$x[which.max(boxcox_mile$y)]

# Box-cox Plot for the mile variable
boxcoxPlotMile <- ggplot(data = boxcox_mileDF, mapping = aes(x = lambda, y = ll)) +
  geom_line(color = "black", linewidth = 0.5) +
  geom_vline(xintercept = maxlabmda_mile, linetype = "dashed", color = "red") +
  geom_hline(yintercept = max(boxcox_mile$y) - qchisq(0.95, df = 1)/2, linetype = "dashed", color = "blu
  labs(title = "Log-likelihood for the Box-Cox transformation",
       x = expression(lambda),
       y = "Log-likelihood for mileage") +
  annotate("text", label = paste0("lambda ==", maxlabmda_mile), x = 1, y = -30000, parse = T) +
  theme_bw() + theme(title = element_text(size=8))

# Create a column of logprice in dat_bmw2
dat_bmw2 <- dat_bmw2 %>% mutate(logprice = log(price))

# Scatter plot of logprice vs. mileage with highlighted outliers
scatterPlotLogPriceMileage <- ggplot(data = dat_bmw2, mapping = aes(x = mileage, y = logprice)) +
  geom_point(pch=19, cex=0.3) +
  geom_point(data = subset(dat_bmw2, mileage > 500000), mapping = aes(x = mileage, y = logprice), pch=19
  geom_point(data = subset(dat_bmw2, logprice > log(100000)), mapping = aes(x = mileage, y = logprice),
  labs(title = "Scatter plot between log(price) and mileage") +
  theme_bw() + theme(title = element_text(size=8))

plot_grid(scatterPlotPriceMileage, boxcoxPlotPrice, boxcoxPlotMile, scatterPlotLogPriceMileage, align =
```



Outliers

```
modelDF <- dat_bmw2
### Assigning NA to car with mileage > 50,0000 miles
modelDF[modelDF$mileage > 500000, "mileage"] <- NA
### Assigning NA to car with price > 100000 usd
modelDF[modelDF$price > 100000, "price"] <- NA
modelDF <- modelDF[complete.cases(modelDF),]
modelDF$logprice <- log(modelDF$price)
```

model fitting We fitted a Ordinary Least Square regression model to predict the natural logrithm of BMW car prices using mileage. The model can be specified as:

$$log(price) = \beta_0 + \beta_1 \times mileage + \epsilon$$

The residuals tell us the difference between

The estimated intercept $\hat{\beta}_0$ of 24600 represents the expected price when mileage is zero, and the slope coefficient $\hat{\beta}_1$ of around -0.062 shows that each additional mileage reduces the price of the vehicle by roughly 6 cents. Both parameter estimates are highly statistically significant with a p-value less than 2×10^{-16}, suggesting that mileage may have a significant relationship with price. The large F-statistic with a p-value, also less than 2×10^{-16} shows that this model is significantly better than the null model (i.e., a model with no predictors). However, the R^2 of 0.1668 indicates that mileage alone explains only about 16.7% of the variation in price, indicating that additional factors likely also influence the price.

```r
mod1 <- lm(logprice ~ mileage, data = modelDF)
summary(mod1)
```

```
##
## Call:
## lm(formula = logprice ~ mileage, data = modelDF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1907 -0.2404  0.0496  0.3229  1.3558
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.017e+01  2.175e-02   467.7   <2e-16 ***
## mileage     -4.746e-06  1.425e-07   -33.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5838 on 4837 degrees of freedom
## Multiple R-squared:  0.1865, Adjusted R-squared:  0.1864
## F-statistic:  1109 on 1 and 4837 DF,  p-value: < 2.2e-16
```