Assignment 9: How much for that car?

Minh Tran

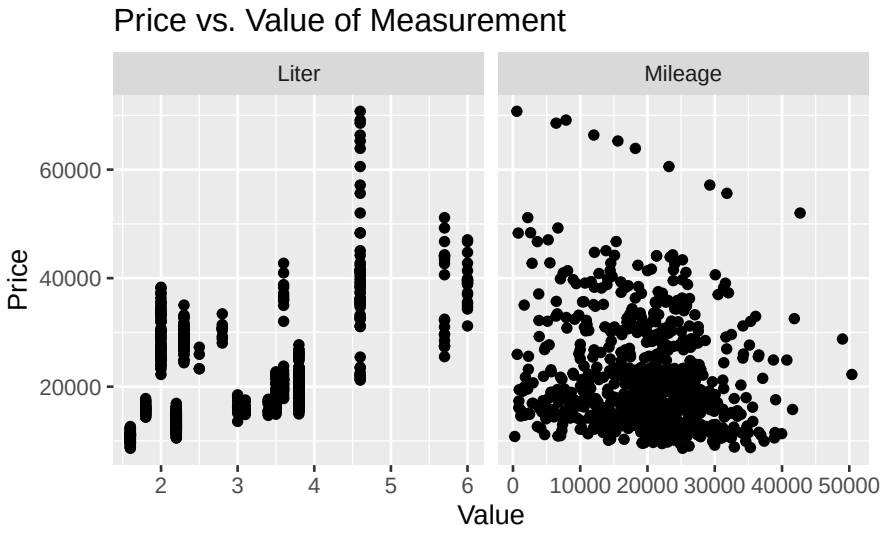
2021-06-07

Exercise 1

i. The other continuous variable is 'Mileage'.

ii.

```
car_prices %>%
  gather(Mileage,Liter, key = "Measurement", value = "Value") %>%
  ggplot() +
  geom_point(mapping = aes(x = Value, y = Price)) +
  facet_wrap(~Measurement, scales = "free_x") +
  labs(title="Price vs. Value of Measurement ")
```



Exercise 2

```
continuous_model <- lm(Price ~ Liter + Mileage, data = car_prices)
```

```
continuous_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	9426.6014688	1095.0777745	8.608157	0.0e+00
Liter	4968.2781155	258.8011436	19.197280	0.0e+00

term	estimate	std.error	statistic	p.value
Mileage	-0.1600285	0.0349084	-4.584237	5.3e-06

```
continuous_model %>%
  glance() %>%
  glimpse()
```

```
## Rows: 1
## Columns: 12
## $ r.squared     <dbl> 0.3291279
## $ adj.r.squared <dbl> 0.3274528
## $ sigma         <dbl> 8106.466
## $ statistic     <dbl> 196.4841
## $ p.value       <dbl> 3.708604e-70
## $ df            <dbl> 2
## $ logLik        <dbl> -8375.659
## $ AIC           <dbl> 16759.32
## $ BIC           <dbl> 16778.08
## $ deviance      <dbl> 52637552164
## $ df.residual   <int> 801
## $ nobs          <int> 804
```

The r-squared value indicates that the model explains about 33% of variation of price is explained by both the Liter and Mileage explanatory variables.

Exercise 3

```
# predict model plane over values
lit <- unique(car_prices$Liter)
mil <- unique(car_prices$Mileage)
grid <- with(car_prices, expand.grid(lit, mil))
d <- setNames(data.frame(grid), c("Liter", "Mileage"))
vals <- predict(continuous_model, newdata = d)

# form surface matrix and give to plotly
m <- matrix(vals, nrow = length(unique(d$Liter)), ncol = length(unique(d$Mileage)))
p <- plot_ly() %>%
  add_markers(
    x = ~car_prices$Mileage,
    y = ~car_prices$Liter,
    z = ~car_prices$Price,
    marker = list(size = 1)
    ) %>%
  add_trace(
    x = ~mil, y = ~lit, z = ~m, type="surface",
    colorscale=list(c(0,1), c("yellow","yellow")),
    showscale = FALSE
```

```
    ) %>%
  layout(
    scene = list(
      xaxis = list(title = "mileage"),
      yaxis = list(title = "liters"),
      zaxis = list(title = "price")
    )
  )
if (!is_pdf) {p}
```

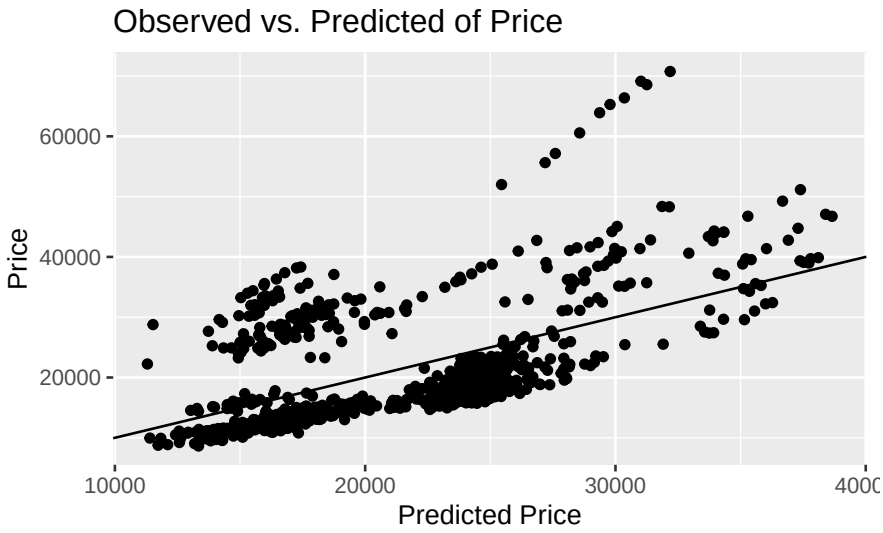
It seems as if the model does not fit the data that well. It is kind of hard to determine if the model meets the three assumptions. It is easier to analyze a 2d model versus a 3d model.

Exercise 4

```
continuous_df <- car_prices %>%
  add_predictions(continuous_model) %>%
  add_residuals(continuous_model)
```

Exercise 5

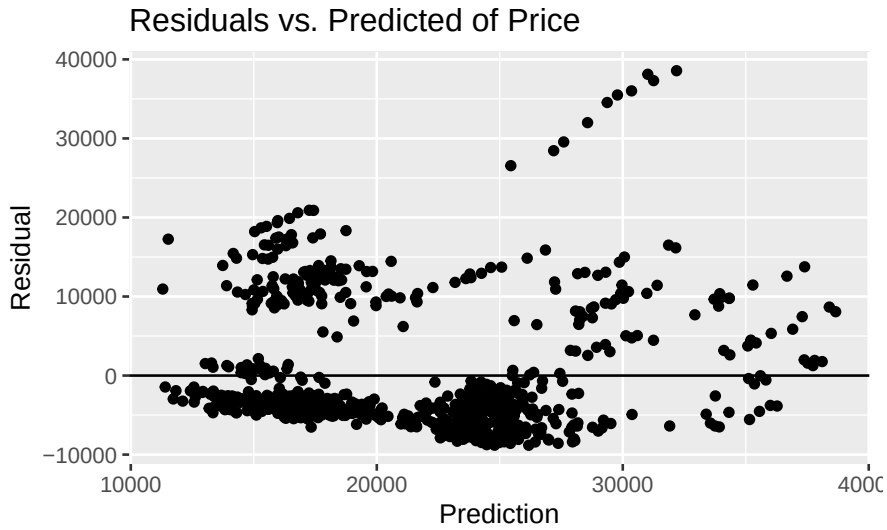
```
continuous_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = Price)) +
  geom_abline(slope = 1, intercept = 0) +
  labs(title = 'Observed vs. Predicted of Price',
       x = 'Predicted Price')
```



This plot shows that the model meets the linear model's assumption of linearity.

Exercise 6

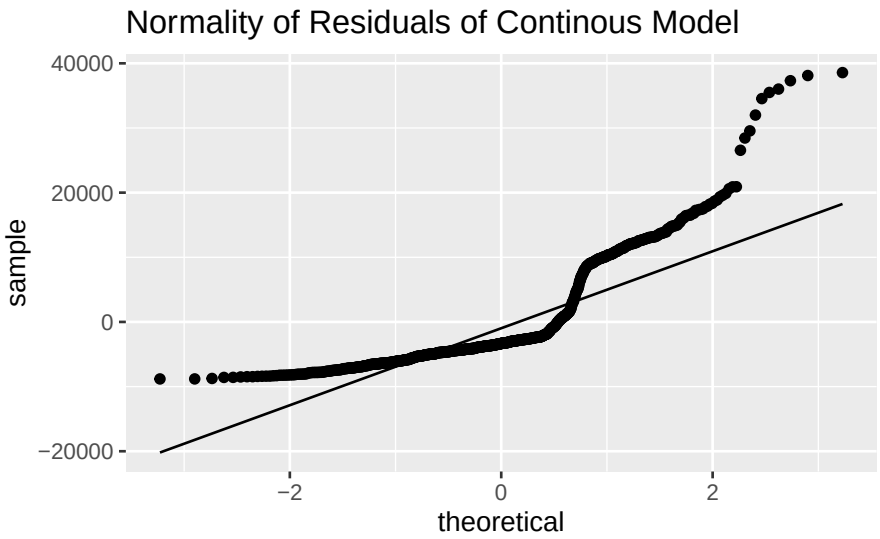
```
continuous_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = resid)) +
  geom_hline(mapping = aes(yintercept = 0)) +
  labs(title = 'Residuals vs. Predicted of Price',
       x = 'Prediction',
       y = 'Residual')
```



This plot shows that the linear model's assumption of constant variability in the residuals is not really met because they are not that centered around zero and shows that the positive residuals have a greater distance.

Exercise 7

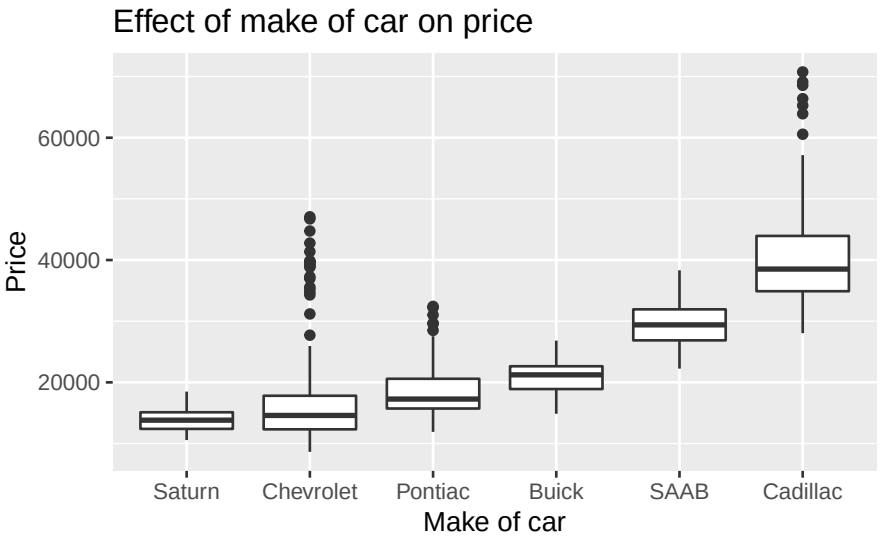
```
continuous_df %>%
  ggplot() +
  geom_qq(aes(sample = resid)) +
  geom_qq_line(aes(sample = resid)) +
  labs(title = 'Normality of Residuals of Continous Model')
```



This plot shows that the linear model's assumption that the residuals are nearly normally distributed is violated since the tails are heavily skewed.

Exercise 8

```
car_prices %>%
  ggplot() +
  geom_boxplot(aes(x = reorder(Make, Price, FUN=median), y = Price)) +
  labs(x = "Make of car", title = "Effect of make of car on price")
```



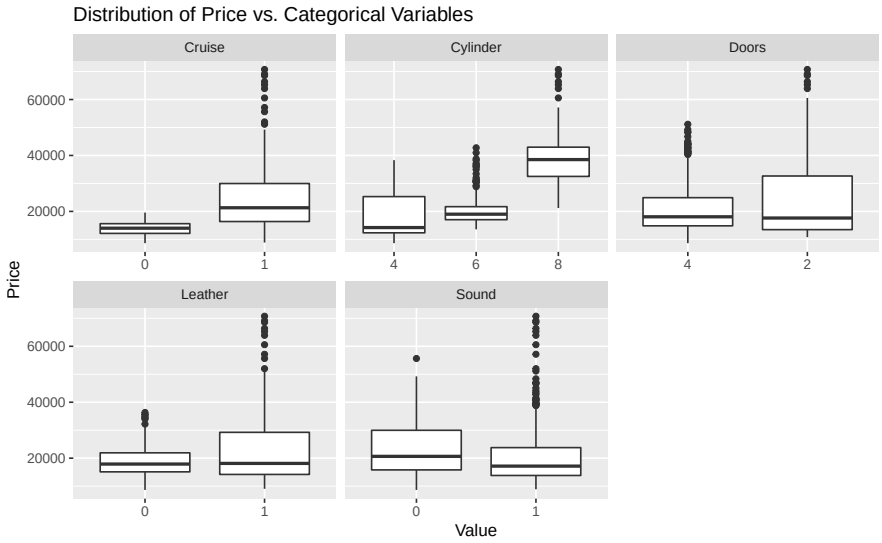
i. ***The make of car that has the lowest median price is Saturn.***

ii. ***Cadillac has the greatest interquartile range of prices.***

iii. ***Cadillac, Chevrolet, and Pontiac have outliers.***

Exercise 9

```
car_prices %>%
  select(-Liter) %>%
  gather(Cylinder:Leather,key="original_column", value="value") %>%
  ggplot() +
  geom_boxplot(aes(x = reorder(value, Price, FUN=median), y = Price)) +
  facet_wrap(~original_column, scales = "free_x") +
  labs(title = "Distribution of Price vs. Categorical Variables",
       x = 'Value')
```



Exercise 10

```
cars_factor_df <- car_prices %>%
  mutate(Cylinder = as.factor(Cylinder))
```

i.

```
mixed_model <- lm(Price ~ Mileage + Liter + Cylinder + Make + Type, data = cars_factor_df)
```

ii.

```
mixed_model %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	1.885018e+04	892.4119413	21.122738	0.0000000
Mileage	-1.861764e-01	0.0106433	-17.492387	0.0000000
Liter	5.697442e+03	342.7322419	16.623596	0.0000000
Cylinder6	-3.312544e+03	619.9683651	-5.343086	0.0000001
Cylinder8	-3.672597e+03	1246.2162662	-2.946998	0.0033032
MakeCadillac	1.450444e+04	517.9855224	28.001635	0.0000000
MakeChevrolet	-2.270807e+03	355.9736337	-6.379145	0.0000000

term	estimate	std.error	statistic	p.value
MakePontiac	-2.355468e+03	363.9063301	-6.472731	0.0000000
MakeSAAB	9.905074e+03	450.2011112	22.001443	0.0000000
MakeSaturn	-2.090266e+03	470.8305609	-4.439529	0.0000103
TypeCoupe	-1.163869e+04	464.7055454	-25.045297	0.0000000
TypeHatchback	-1.172638e+04	545.3936364	-21.500769	0.0000000
TypeSedan	-1.178618e+04	411.1021489	-28.669707	0.0000000
TypeWagon	-8.156551e+03	500.6379995	-16.292312	0.0000000

iii.

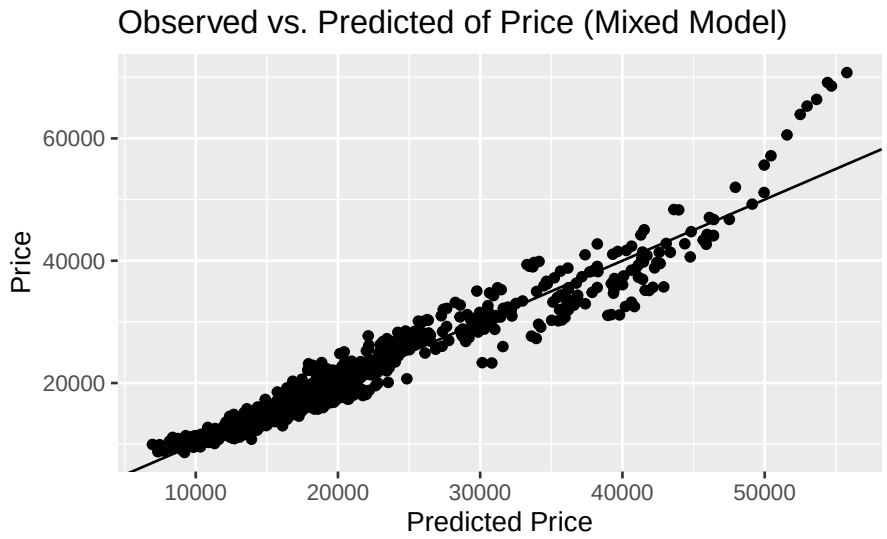
```
mixed_model %>%
  glance() %>%
  glimpse()
```

```
## Rows: 1
## Columns: 12
## $ r.squared     <dbl> 0.9389165
## $ adj.r.squared <dbl> 0.9379113
## $ sigma         <dbl> 2463.068
## $ statistic     <dbl> 934.0858
## $ p.value       <dbl> 0
## $ df            <dbl> 13
## $ logLik        <dbl> -7412.332
## $ AIC           <dbl> 14854.66
## $ BIC           <dbl> 14925.01
## $ deviance      <dbl> 4792696059
## $ df.residual   <int> 790
## $ nobs          <int> 804
```

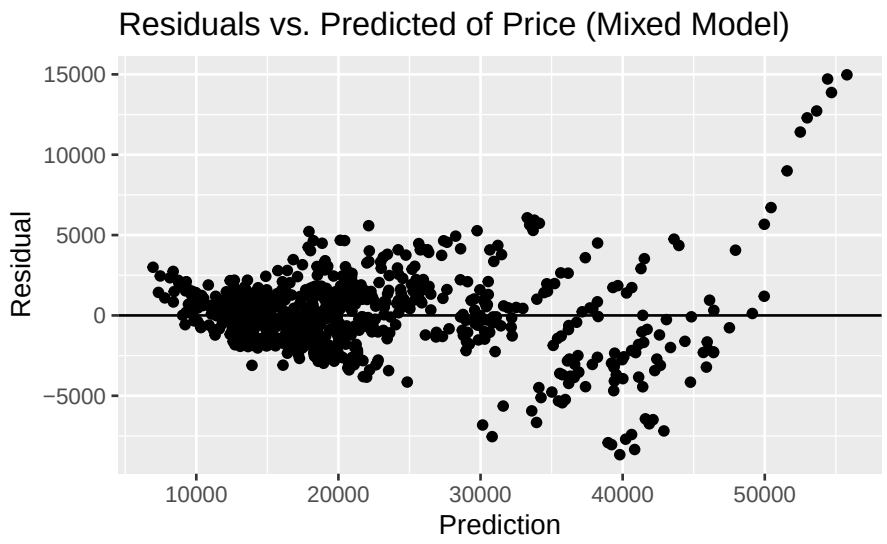
Exercise 11

```
mixed_df <- cars_factor_df %>%
  add_predictions(mixed_model) %>%
  add_residuals(mixed_model)
```

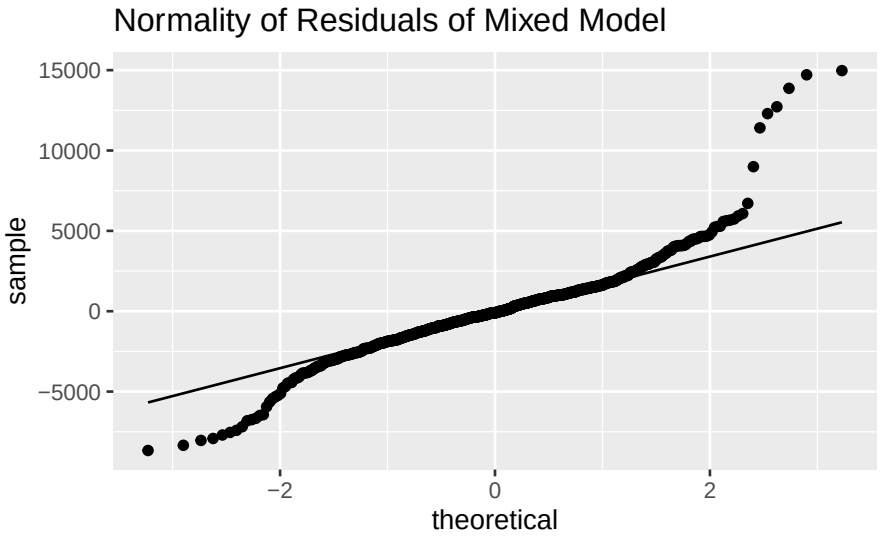
```
mixed_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = Price)) +
  geom_abline(slope = 1, intercept = 0) +
  labs(title = 'Observed vs. Predicted of Price (Mixed Model)',
       x = 'Predicted Price')
```




```
mixed_df %>%
  ggplot() +
  geom_point(mapping = aes(x = pred, y = resid)) +
  geom_hline(mapping = aes(yintercept = 0)) +
  labs(title = 'Residuals vs. Predicted of Price (Mixed Model)',
       x = 'Prediction',
       y = 'Residual')
```




```
mixed_df %>%
  ggplot() +
  geom_qq(aes(sample = resid)) +
  geom_qq_line(aes(sample = resid)) +
  labs(title = 'Normality of Residuals of Mixed Model')
```



Exercise 12

i. ***The higher r-squared value in the mixed model indicates that the mixed model better explains the variation in price than the continous model. The mixed model meets the linear model's three assumptions better than the continous model. The continous model violated at least 2 of the assumptions.***

ii. ***If I was picking a car, I would choose to use the mixed model since it better explains the variation of price (93%). Not only that, but the continous model is not a good model since it violated at least two of the three linear model's assumptions. On the other hand, the mixed model better meets the three assumptions. The mixed model clearly meets the linearity assumption, it meets the constant variability assumption since the points are centered around the horizontal line of zero, and meets the the normality of residuals since the points line up with the line.***