# Medical Cost Insurance Dependencies in the USA

Daniel Osheroff, Bedant Lohani

4/20/2021

Contributions: We split the contributions using our goals written in the proposal.

Daniel Osheroff: Goals 2 and 3:

- Use the best potential predictor/s variable/s to build the best possible linear regression model with charges as the response variables.
- Validate the model assumptions and measure the model quality by R^2 and MSPE.

Bedant Lohani: Goals 1 and 4:

- Visualise and comment on the relationship between charges/medical cost billed by health insurance and other qualitative as well as quantitative data attributes/variables in the dataset.
- Predict the average cost of insurance for a particular group of people based on the best potential predictor variables by giving a reasonable confidence interval.

Introduction:

The dataset we are going to use in this project incorporates statistics about medical insurance costs for people in the United States. The people in the data have 7 attribute groups based on sex, age, bmi, number of dependents/children, smoker/non-smoker, region, and charges, but our main objective is to find out if there exists a predictor that can be used to predict the insurance cost/charge and if there exists a linear relation, find out the best possible model to represent it.

We also intend to use the dataset to visualise and summarise the relationship between charges/medical cost billed by health insurance and other qualitative as well as quantitative data attributes in the dataset in the process.

We believe that it is useful to find out what the best predictor of the medical cost/charges is so that necessary preparations or precautions can be taken to reduce the cost by an individual in the USA.

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

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(broom)
library(mosaic)
```

```
## Registered S3 method overwritten by 'mosaic':
##   method                           from
```

```
##   fortify.SpatialPolygonsDataFrame ggplot2
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
## Attaching package: 'mosaic'
## The following object is masked from 'package:Matrix':
##
##     mean
## The following objects are masked from 'package:dplyr':
##
##     count, do, tally
## The following object is masked from 'package:ggplot2':
##
##     stat
## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median, prop.test,
##     quantile, sd, t.test, var
## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
library(infer)
```

```
##
## Attaching package: 'infer'
## The following objects are masked from 'package:mosaic':
##
##     prop_test, t_test
```

```
#read insurance data file
insurance_data = read.csv("insurance.csv")

#shorten name
insurance = insurance_data

#examine data
head(insurance)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

```
dim(insurance)
```

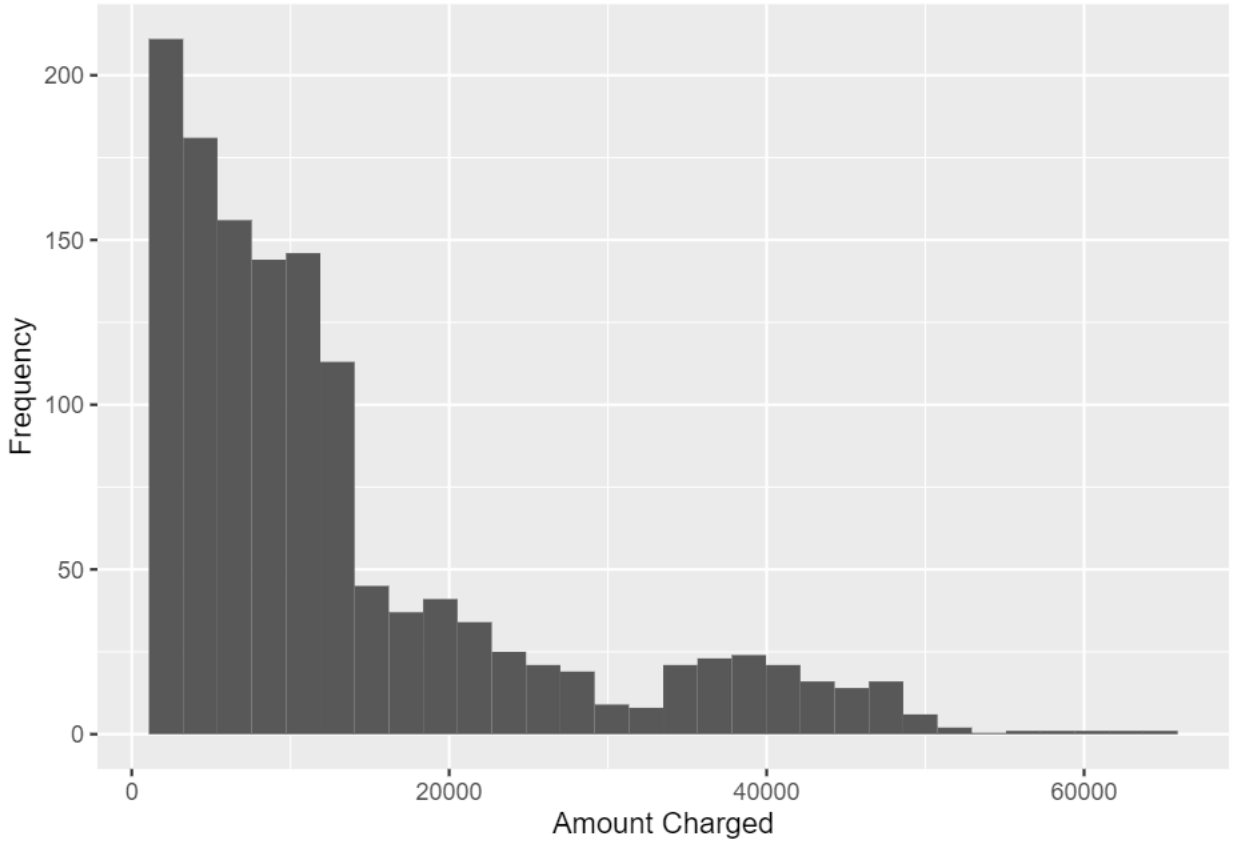
```
## [1] 1338    7
```

```
names(insurance)
```

```
## [1] "age"      "sex"      "bmi"      "children" "smoker"   "region"   "charges"
```

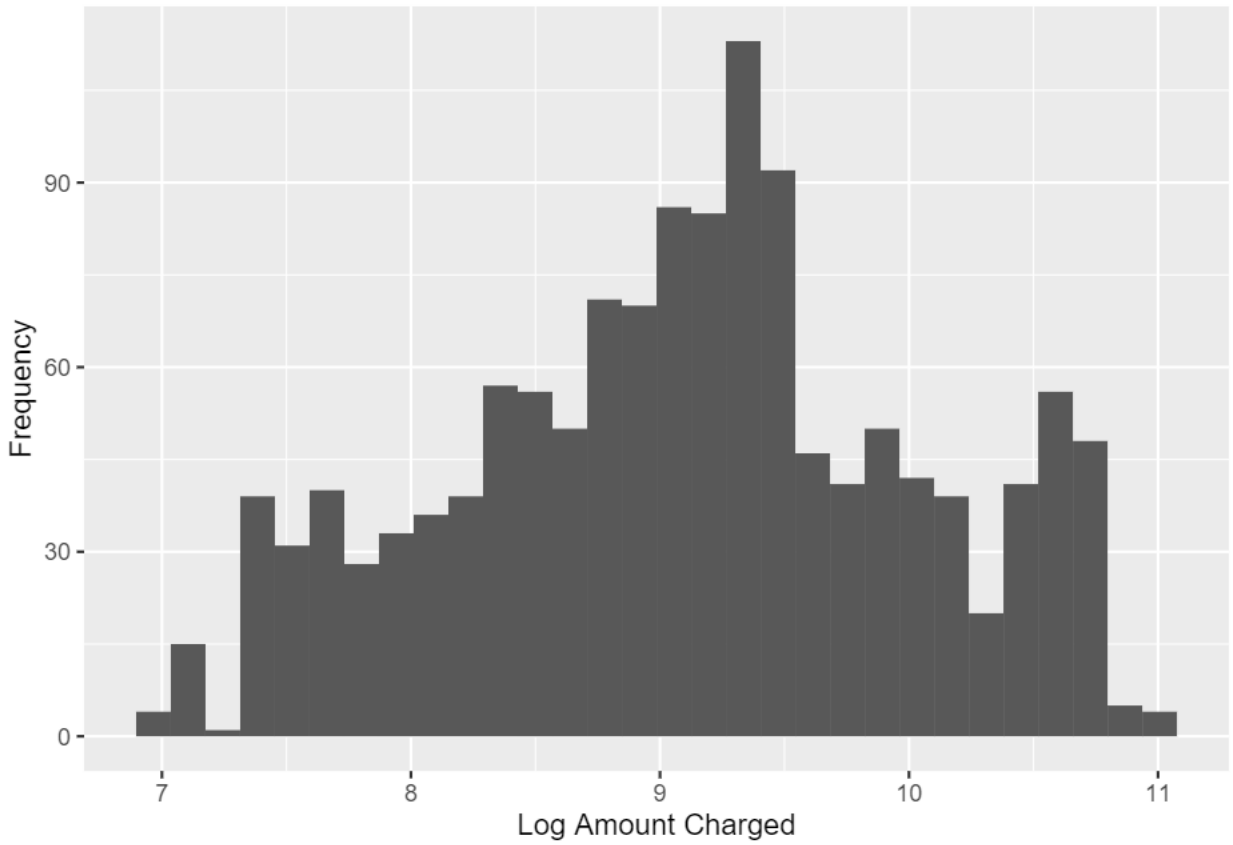
```
#create histogram of charges to understand overall distribution
ggplot(insurance, aes(x = charges)) +
  geom_histogram() +
  labs(x = "Amount Charged", y = "Frequency")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

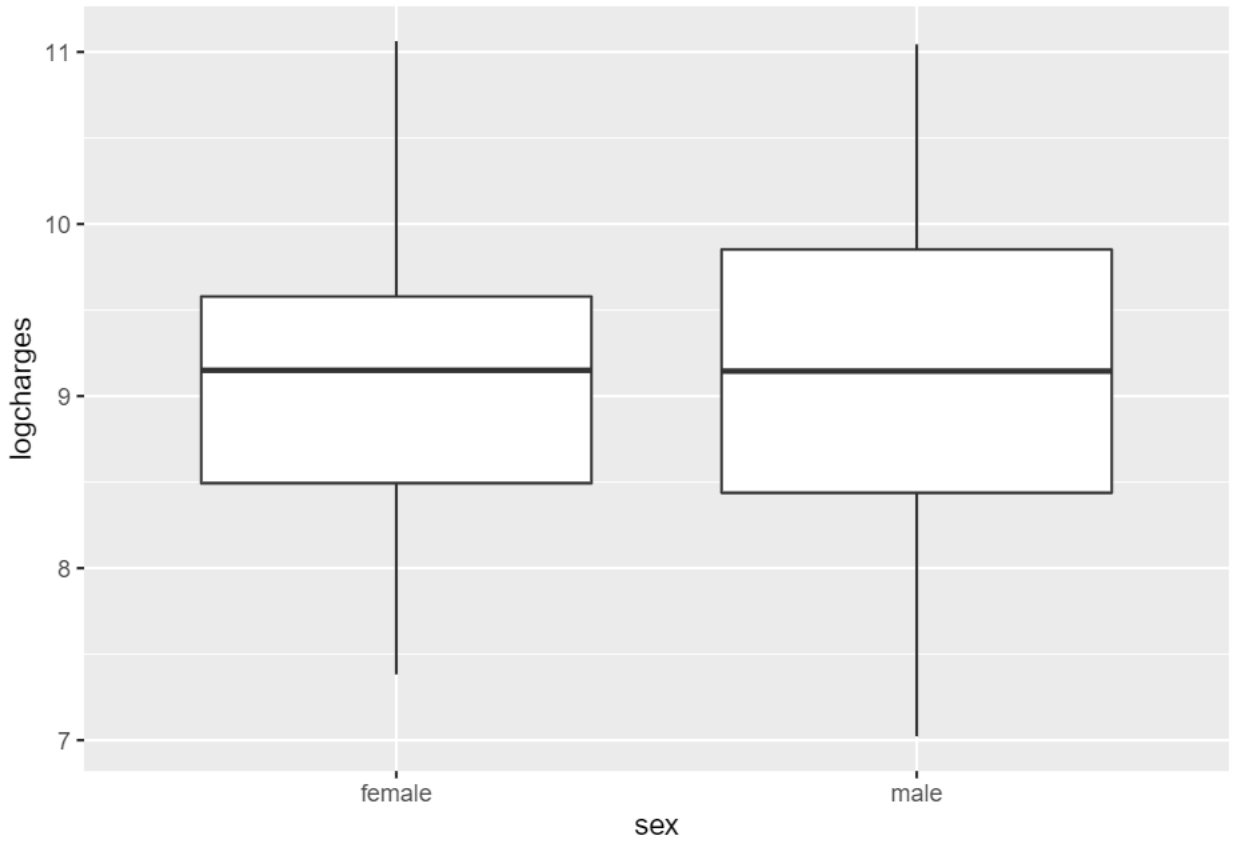



```
# The distribution is clearly heavily skewed to the right, and is unimodal
# with what appears to be a small number of upper outliers
# to remove some of this skew, we will take the logarithm of charges

#mutate the data to take logarithms
insurance = insurance %>%
  mutate(logcharges = log(charges))

ggplot(insurance, aes(x = logcharges)) +
  geom_histogram() +
  labs(x = "Log Amount Charged", y = "Frequency")
```

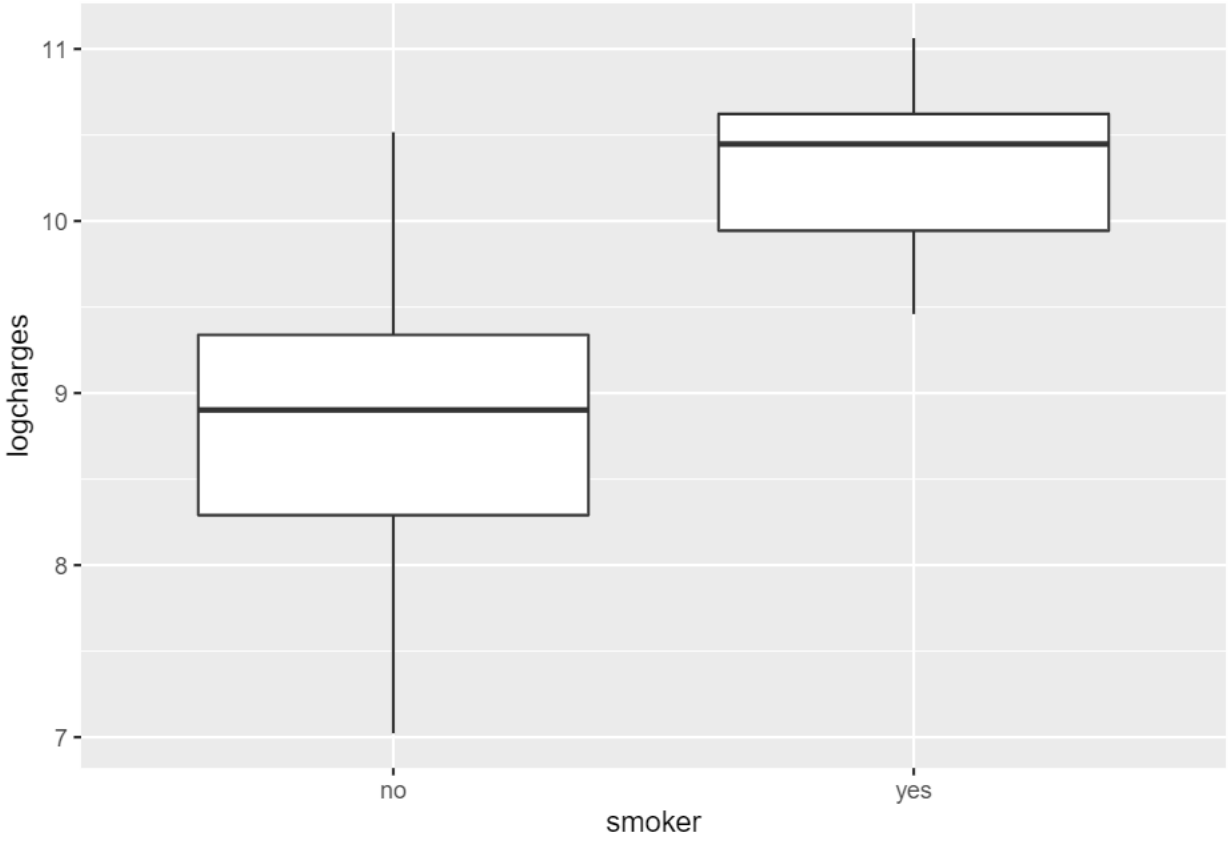
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# One can now see that the data looks far more normal and easy to work with

# Create boxplots for charges v categorical variables (sex, smoker)

ggplot(insurance, aes(y = logcharges, x = sex)) +
  geom_boxplot()
```

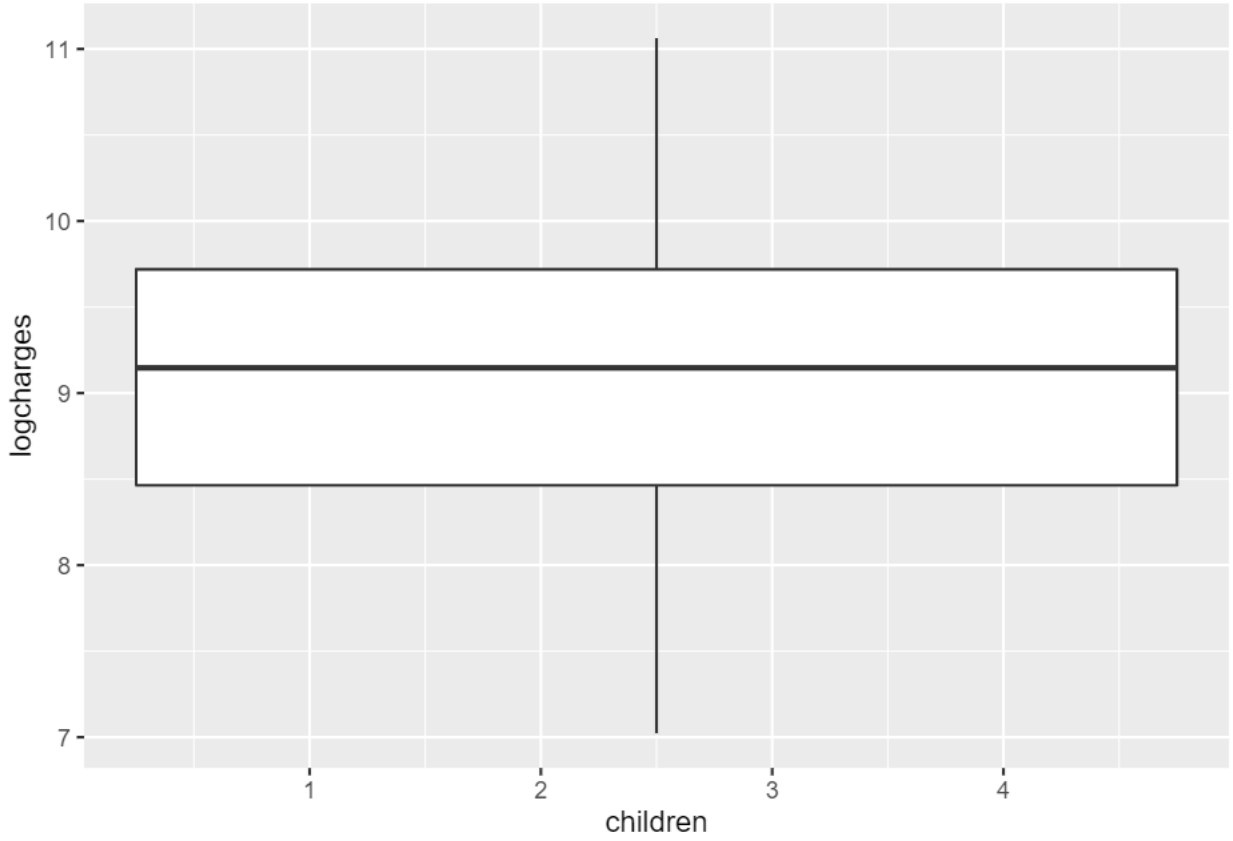



```
# Sex does not seem to impact insurance charges, although the variance for
# males is slightly larger

ggplot(insurance, aes(y = logcharges, x = smoker)) +
  geom_boxplot()
```

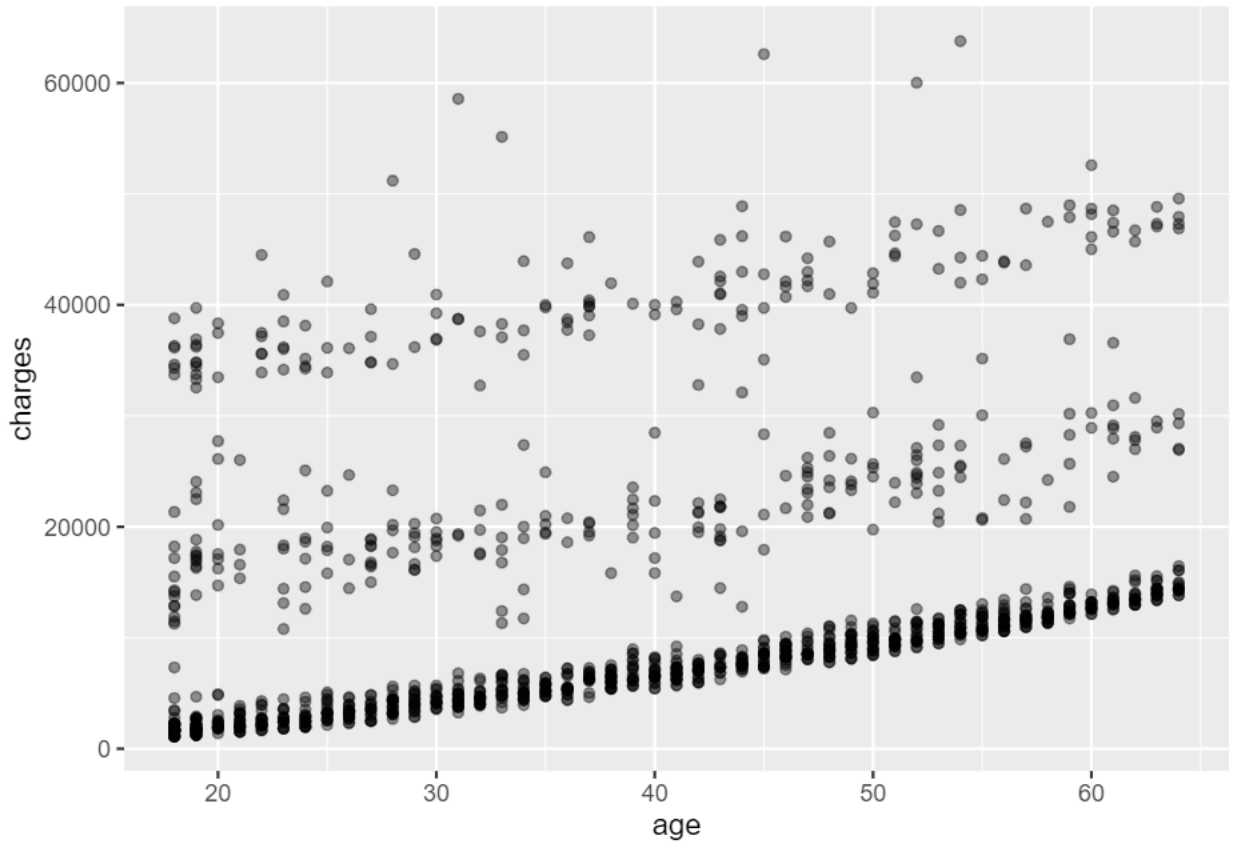



```
# There is a clear relationship that smokers incur larger insurance charges

ggplot(insurance, aes(y = logcharges, x = children)) +
  geom_boxplot()
```

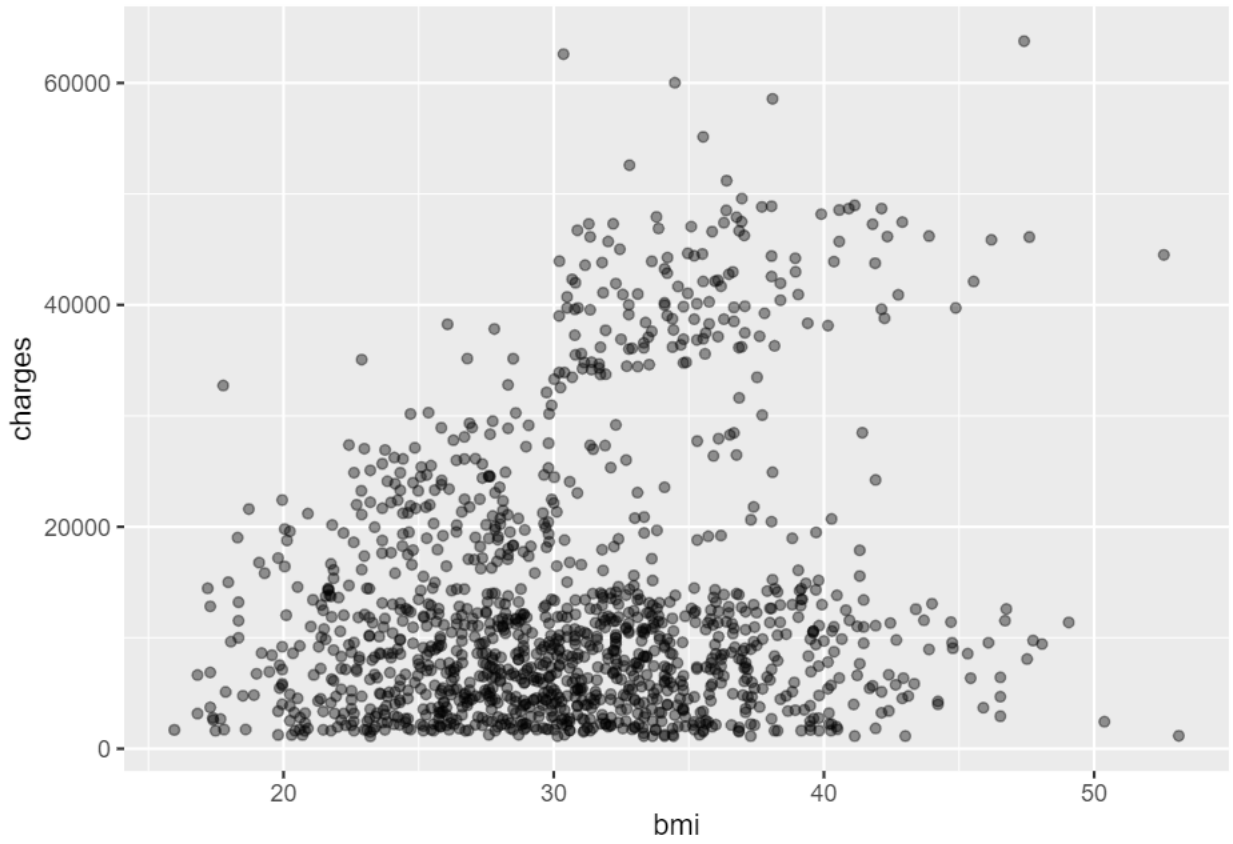
```
## Warning: Continuous x aesthetic -- did you forget aes(group=...)?
```



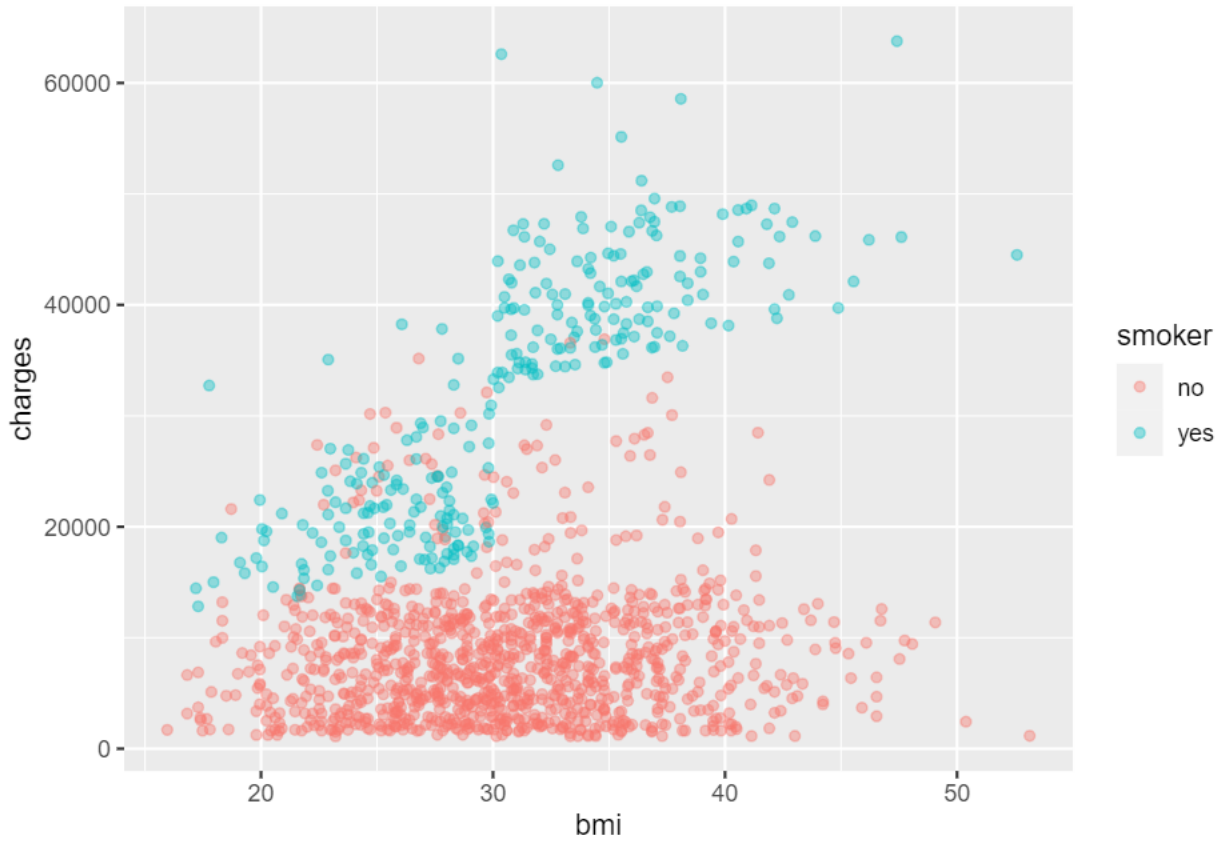

```
# Create scatterplots of charges v age, charges v bmi, charges v children

ggplot(insurance, aes(y = charges, x = age)) +
  geom_point(alpha = 0.4)
```



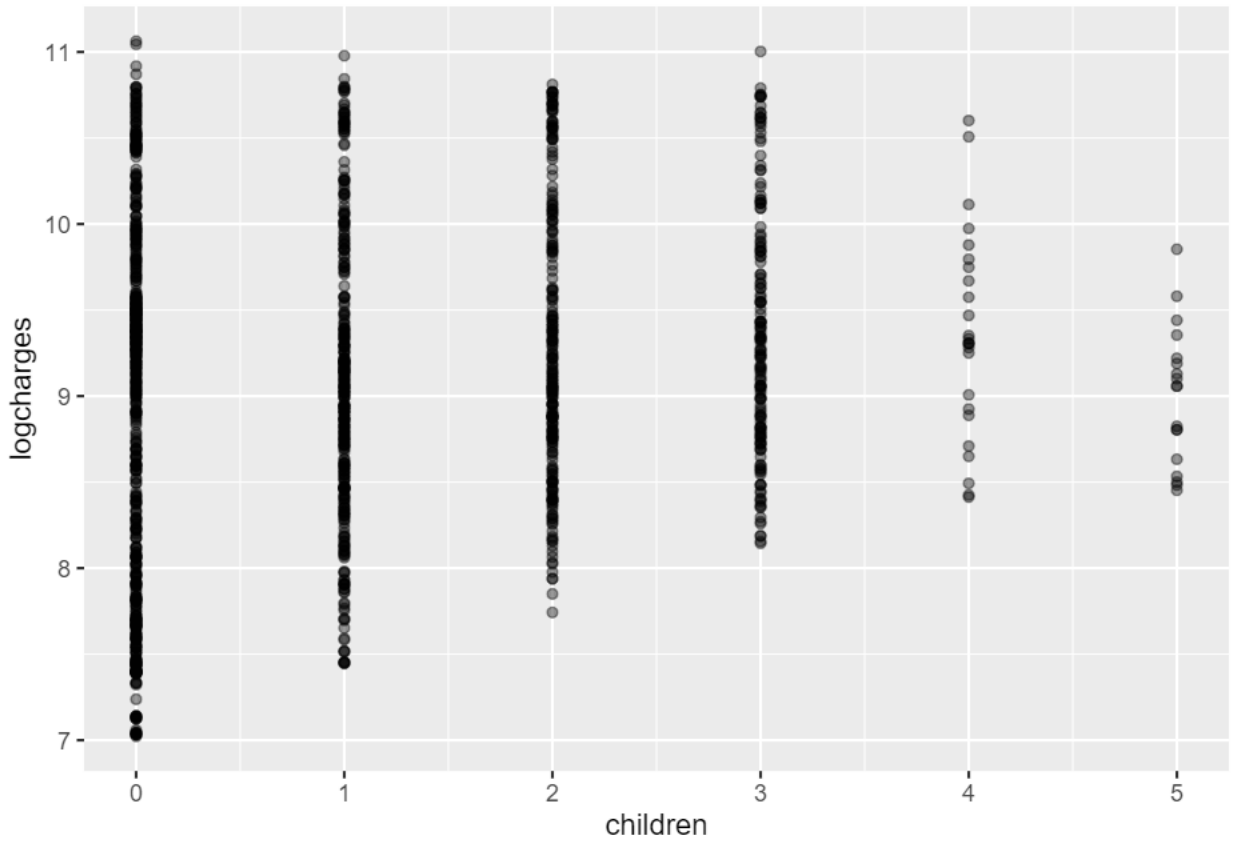

```
# There appears to be an extremely consistent linear relationship between
# insurance charges and age, possibly because older individuals are at
# higher risk of incurring medical cost.

ggplot(insurance, aes(y = charges, x = bmi))+
  geom_point(alpha = 0.4)
```




```
# Interestingly, there appears to be little overall correlation between
# charges and BMI, although there do seem to be 2 distinct groups
# one with a positive linear relationship and one with no relationship

ggplot(insurance, aes(y = charges, x = bmi, color = smoker))+
  geom_point(alpha = 0.4)
```




```
# Controlling for smoking, one sees that although there is nor relationship
# between charges and BMI for nonsmokers, there is a positive correlation between
# charges and BMI for smokers. There is also a jump at bmi = 30 for smokers
# that is not there for nonsmokers, that for the most part seems to double
# the overall cost

ggplot(insurance, aes(y = logcharges, x = children)) +
  geom_point(alpha = 0.4)
```




```
# Although variance increases as number of children decreases, there does not
# appear to be a linear correlation between charges and number of children
```

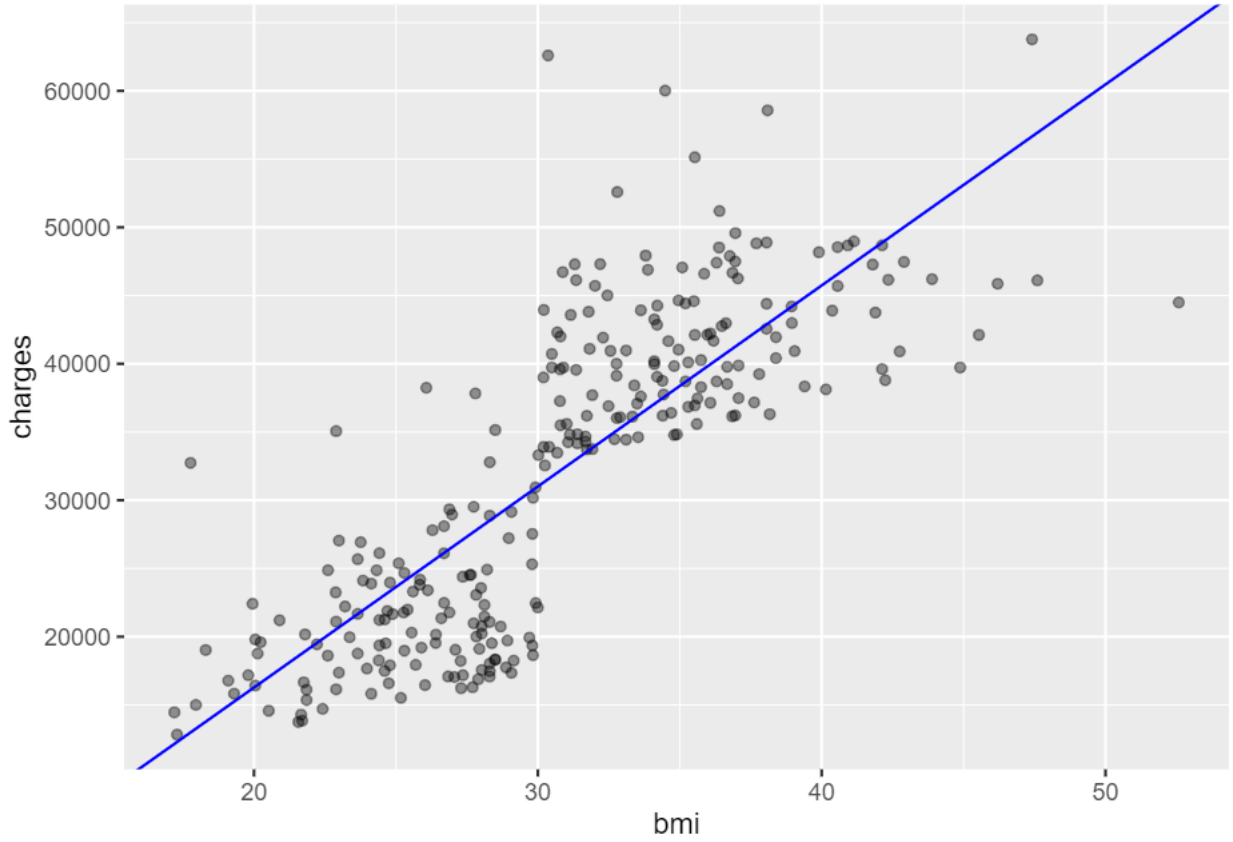
Now that the data has been analyzed and visualized, we will create models based off of the approximately linear relationships viewed. This includes charges v BMI for smokers and charges v age. So that the models may be better compared and combined, we will only use data from smokers for the charges v age model as well.

```
# Subset the data to be smokers only
insurance_smokers = insurance %>%
  filter(smoker == "yes")

# Create model of bmi v charges and print summary
bmi_model = lm(charges ~ bmi, data = insurance_smokers)
summary(bmi_model)
```

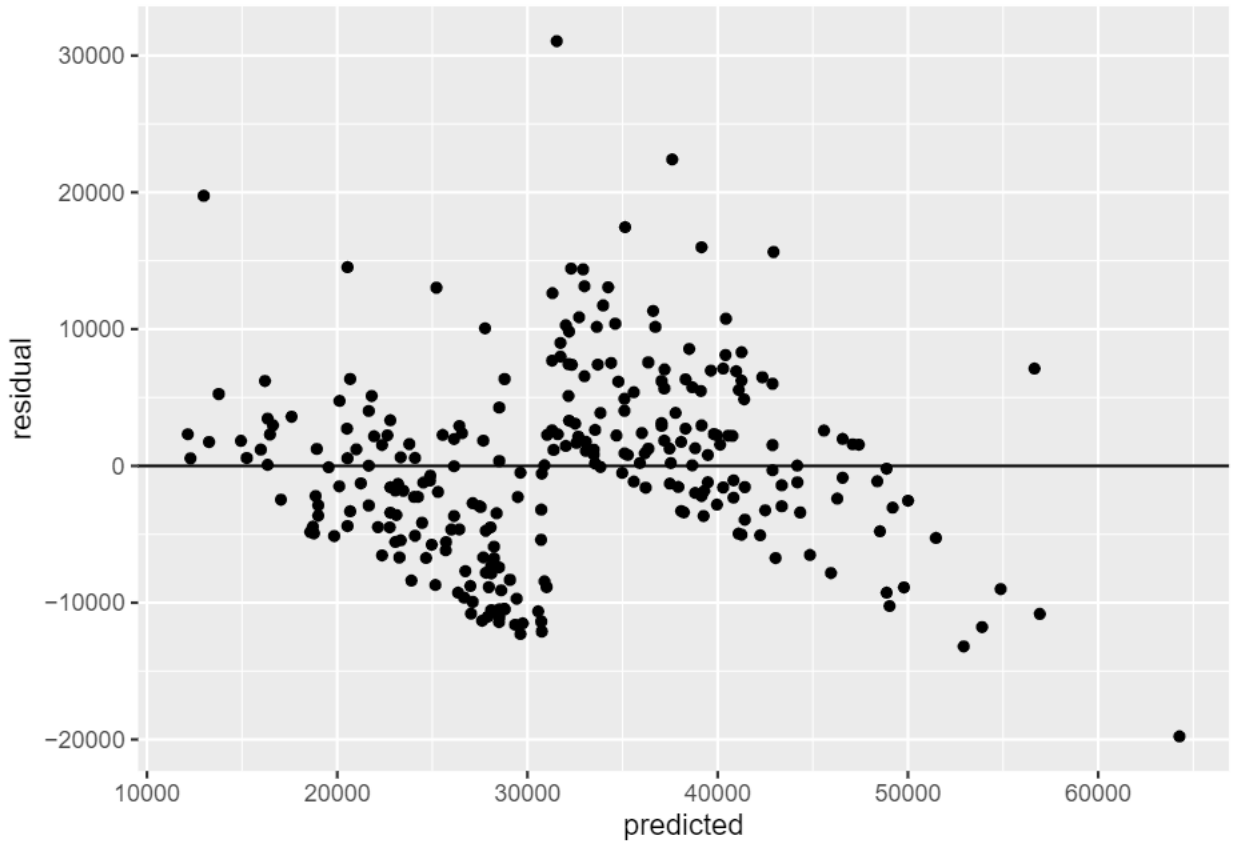
```
##
## Call:
## lm(formula = charges ~ bmi, data = insurance_smokers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19768.0  -4487.9     34.4   3263.9  31055.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13186.58    2052.88  -6.423 5.93e-10 ***
## bmi           1473.11      65.48  22.496  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6837 on 272 degrees of freedom
## Multiple R-squared:  0.6504, Adjusted R-squared:  0.6491
## F-statistic: 506.1 on 1 and 272 DF,  p-value: < 2.2e-16
```

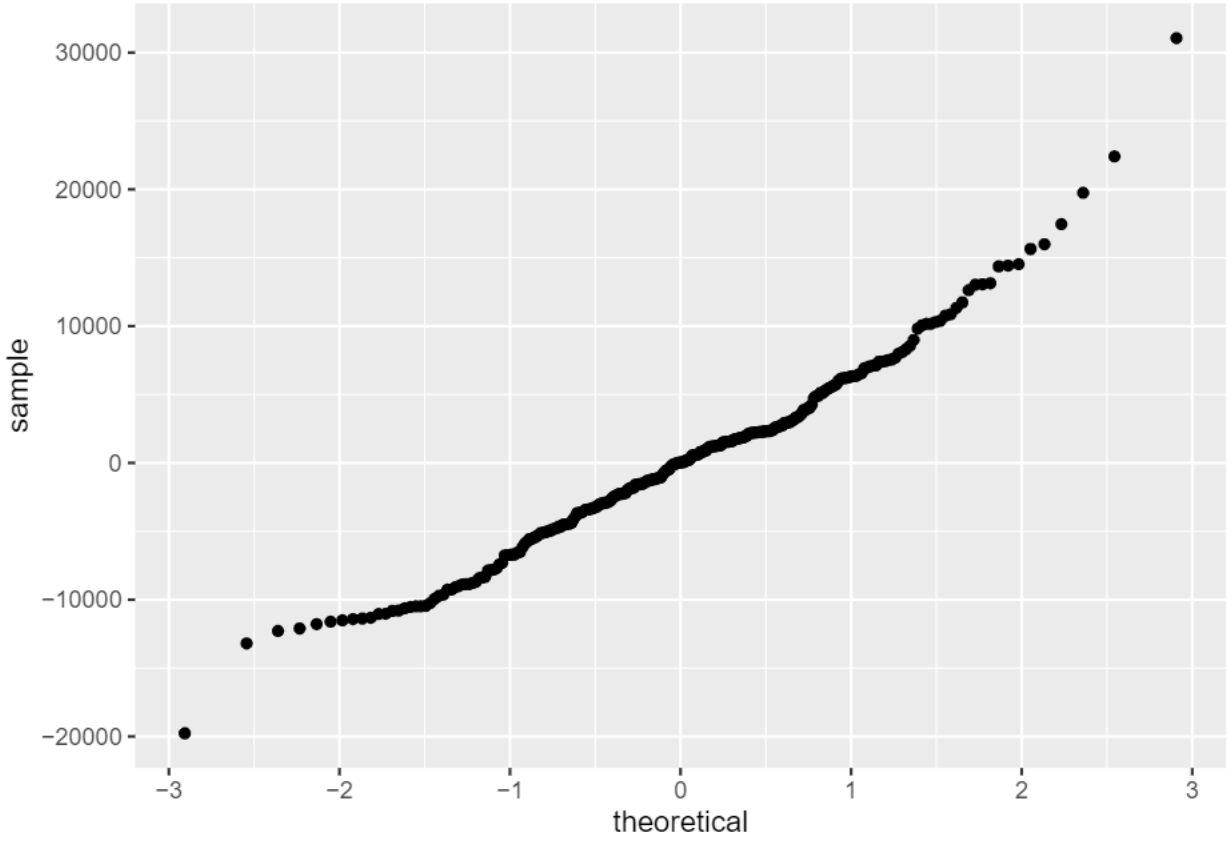
```r
# Assumption 1 is met because it is reasonable to assume one person's bmi and
# insurance charge is independent from another's

# Draw the regression line
ggplot(insurance_smokers, aes(y = charges, x = bmi)) +
  geom_point(alpha = 0.4) +
  geom_abline(slope = 1473.11, intercept = -13186, color = "blue")
```




```r
# Create a data frame with actual values, predicted values, and residuals from
# model

bmi_mod_results = data.frame(observed = insurance_smokers$charges,
                             predicted = bmi_model$fitted.values,
                             residual = bmi_model$residuals)

# Plot the residuals versus the predictions
ggplot(bmi_mod_results, aes(y = residual, x = predicted)) +
  geom_point() +
  geom_hline(yintercept = 0)
```




```
# The residual plot does not look great overall. There are clear trends, with the
# residuals decreasing from predcicted = 9.5 - 10.25, a jump at 10.25, and then
# decreasing again from 10.5 - 11.5. As such, this model isn't great for the
# data, due to the jump at bmi = 30


# Create a Q-Q plot of the residuals
ggplot(bmi_mod_results, aes(sample = residual)) +
  geom_qq()
```




```
# Despite the failings of the residual plot, the Q-Q plot actually looks pretty
# good, and is an approximately straight line
```

Next, we segment the data into bmi $> 30$ and bmi $< 30$ to try to remove some of the failings of the previous model, using the same steps

```
# Filter the data by bmi greater than and less then or equal to 30

lowbmi_smokers = insurance_smokers %>%
  filter(bmi <= 30)

highbmi_smokers = insurance_smokers %>%
  filter(bmi > 30)

# Follow the same steps for model creation and assumption checking

lowbmi_model = lm(charges ~ bmi, data = lowbmi_smokers)
summary(lowbmi_model)
```

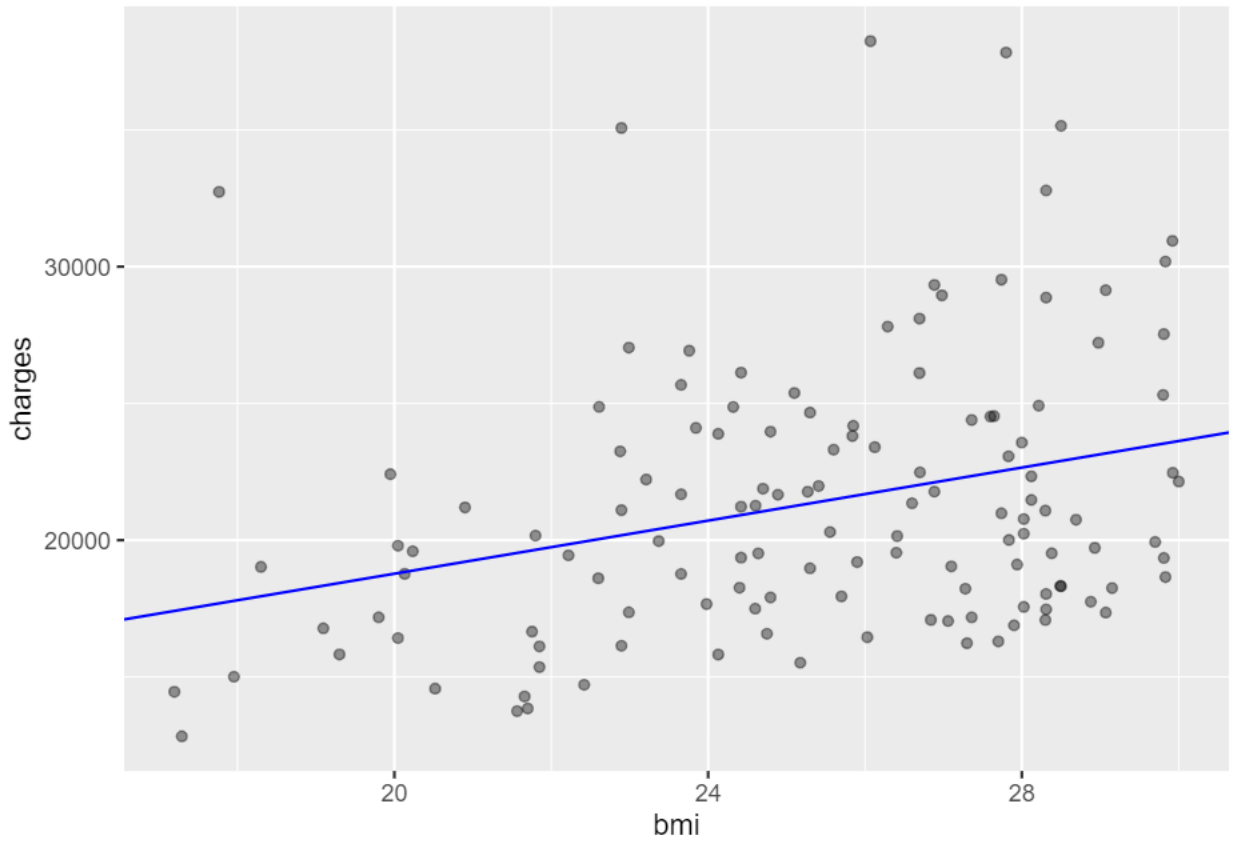
```
##
## Call:
## lm(formula = charges ~ bmi, data = lowbmi_smokers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6211  -3538  -1206   2190  16528
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9071.2     3424.3   2.649 0.009088 **
## bmi            485.1      134.0   3.619 0.000424 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4826 on 128 degrees of freedom
## Multiple R-squared:  0.09283,    Adjusted R-squared:  0.08574
## F-statistic:  13.1 on 1 and 128 DF,  p-value: 0.0004242
```

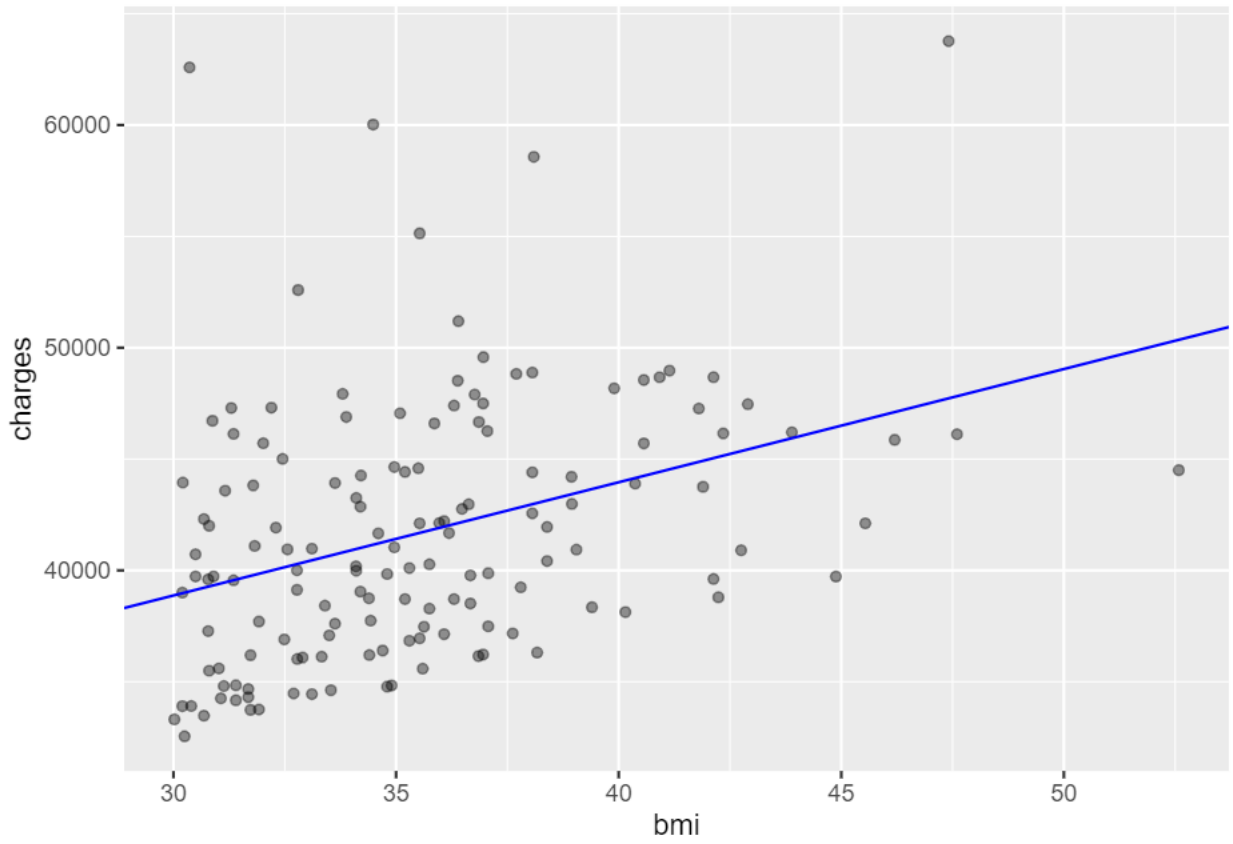
```
highbmi_model = lm(charges ~ bmi, data = highbmi_smokers)
summary(highbmi_model)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = highbmi_smokers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6720.0 -4499.8  -608.4  3213.0 23536.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23617.2     3894.9   6.064 1.14e-08 ***
## bmi            508.5      108.8   4.673 6.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5446 on 142 degrees of freedom
## Multiple R-squared:  0.1333, Adjusted R-squared:  0.1272
## F-statistic: 21.83 on 1 and 142 DF,  p-value: 6.833e-06
```

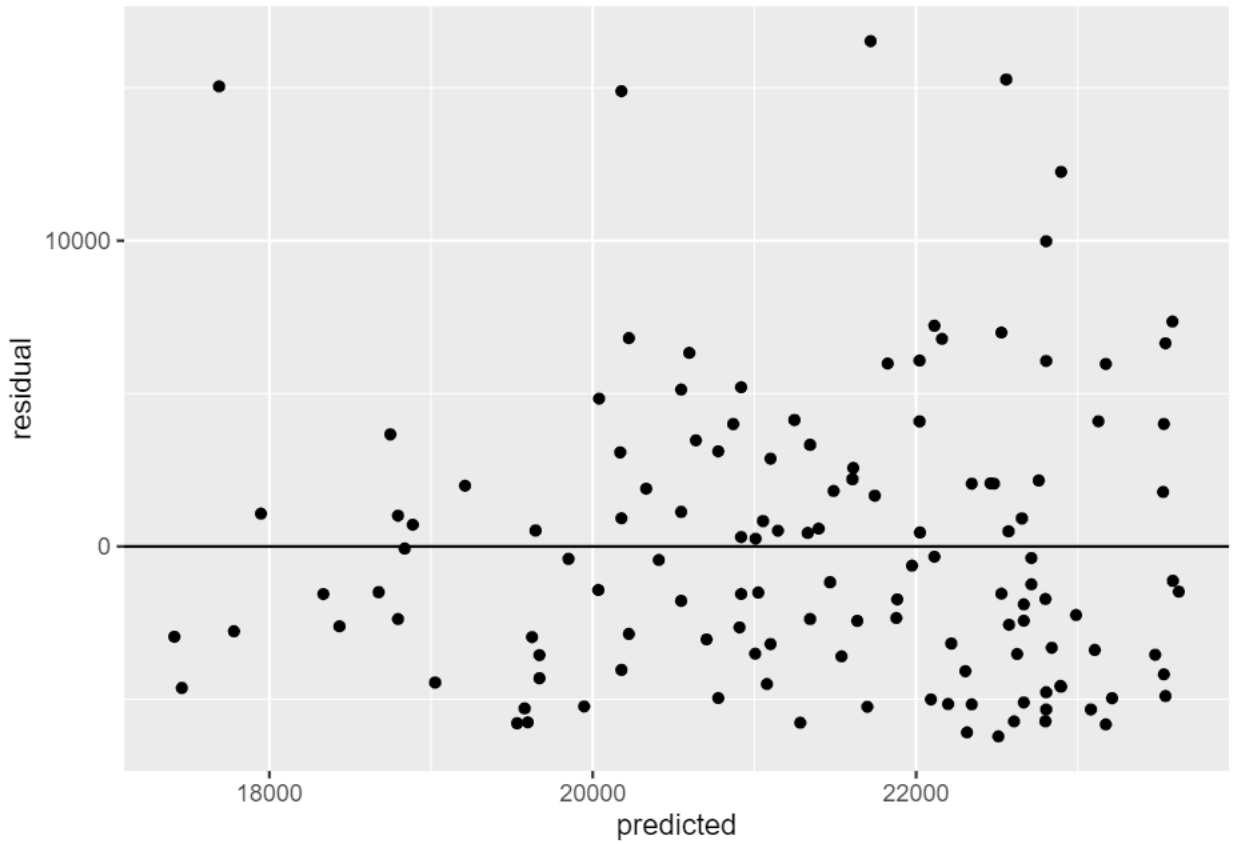
```
ggplot(lowbmi_smokers, aes(y = charges, x = bmi)) +
  geom_point(alpha = 0.4) +
  geom_abline(slope = 485.1, intercept = 9071.2, color = "blue")
```



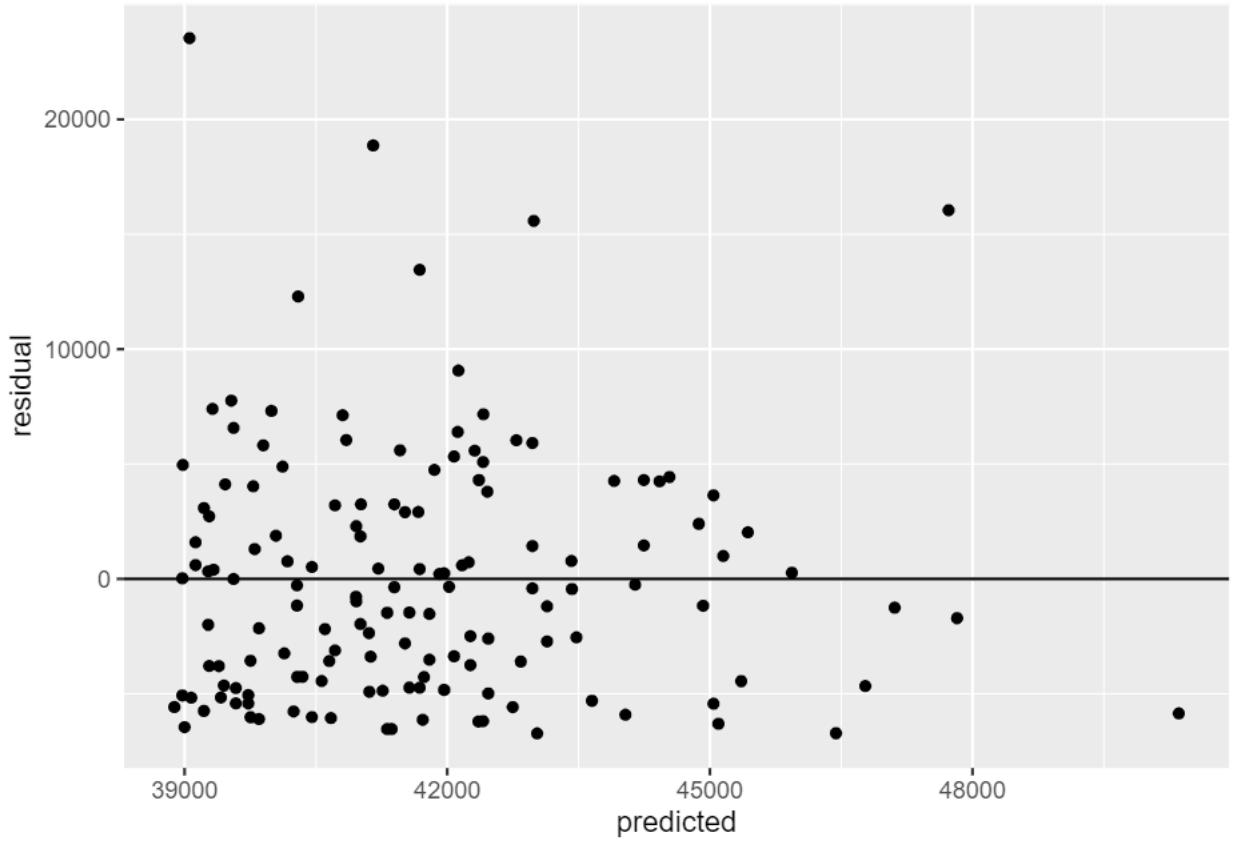

```
ggplot(highbmi_smokers, aes(y = charges, x = bmi)) +
  geom_point(alpha = 0.4) +
  geom_abline(slope = 508.5, intercept = 23617.2, color = "blue")
```



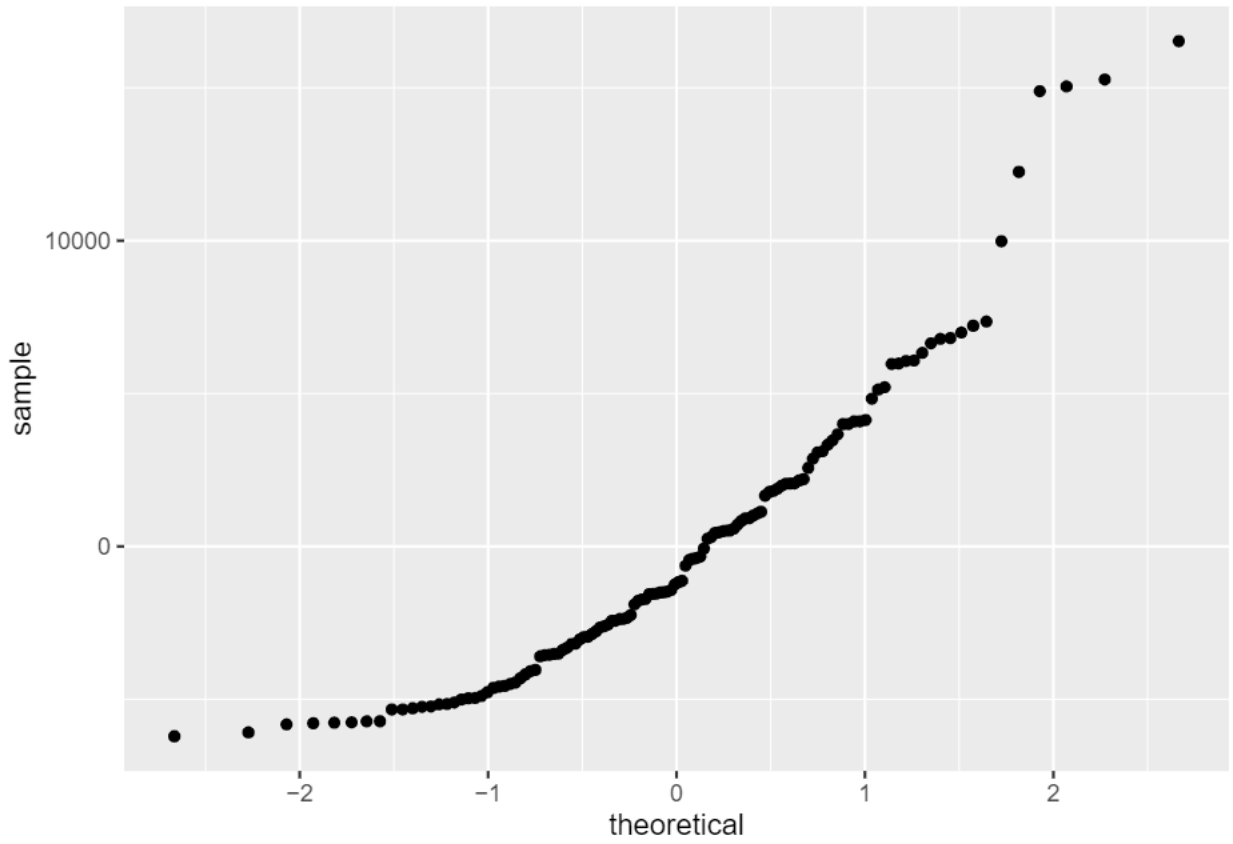

```
lowbmi_mod_results = data.frame(observed = lowbmi_smokers$charges,
                                predicted = lowbmi_model$fitted.values,
                                residual = lowbmi_model$residuals)

highbmi_mod_results = data.frame(observed = highbmi_smokers$charges,
                                 predicted = highbmi_model$fitted.values,
                                 residual = highbmi_model$residuals)

ggplot(lowbmi_mod_results, aes(y = residual, x = predicted)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



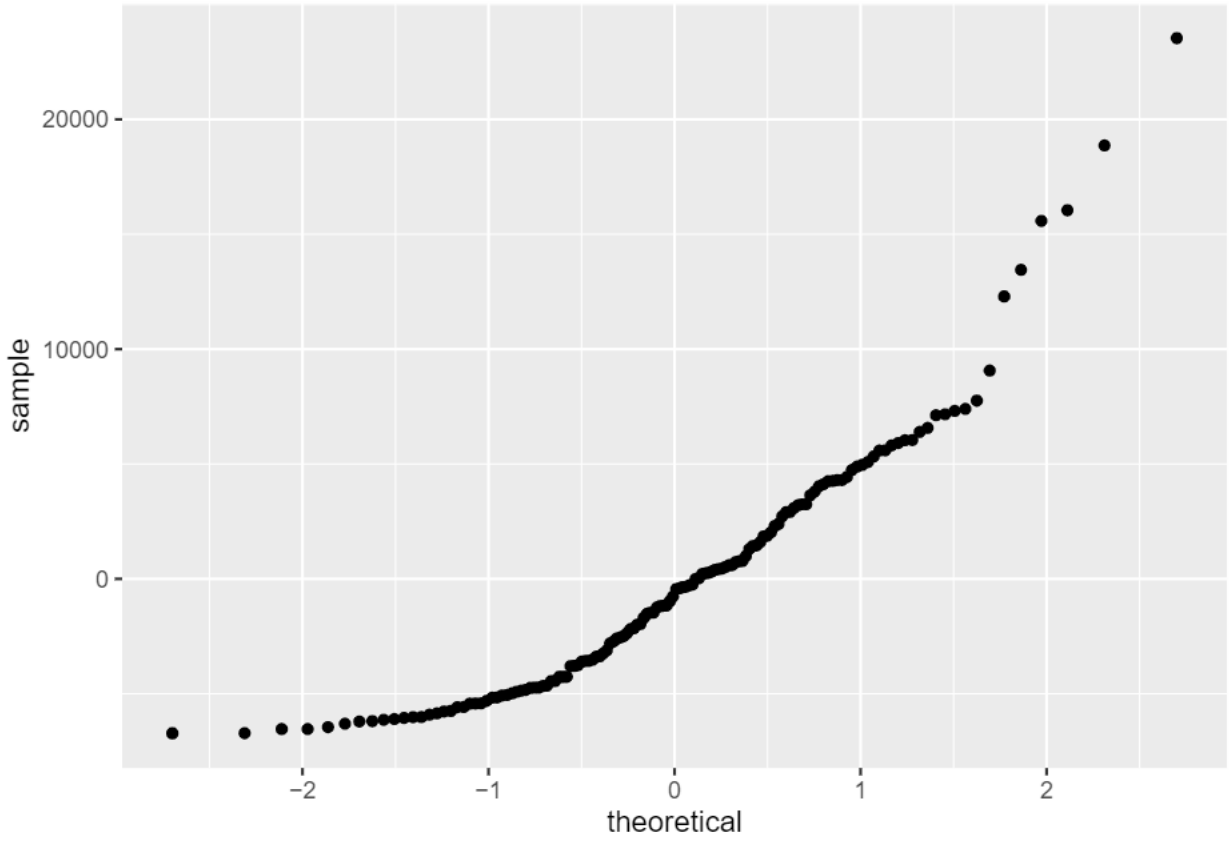

```
ggplot(highbmi_mod_results, aes(y = residual, x = predicted)) +
  geom_point() +
  geom_hline(yintercept = 0)
```




```
# Both of these residual plots look significantly better, with the points
# symmetrically clustered around low values of y with no clear pattern.
# However, there is still failings to the residual plots, as in both, there are
# numerous data points with extremely high residuals, but a lack of the same
# outliers on the lower end

ggplot(lowbmi_mod_results, aes(sample = residual)) +
  geom_qq()
```




```
ggplot(highbmi_mod_results, aes(sample = residual)) +
  geom_qq()
```




```
# The Q-Q plot  looks worse for the more segmented models, perhaps
# because the variance in the overall data tends to grow far greater as
# bmi approaches 30

# From this analysis, it is clear that neither the segmented nor the
# unsegmented (by bmi) data conforms perfectly to a linear model. Quality
# of the models will be assessed later in the project
```

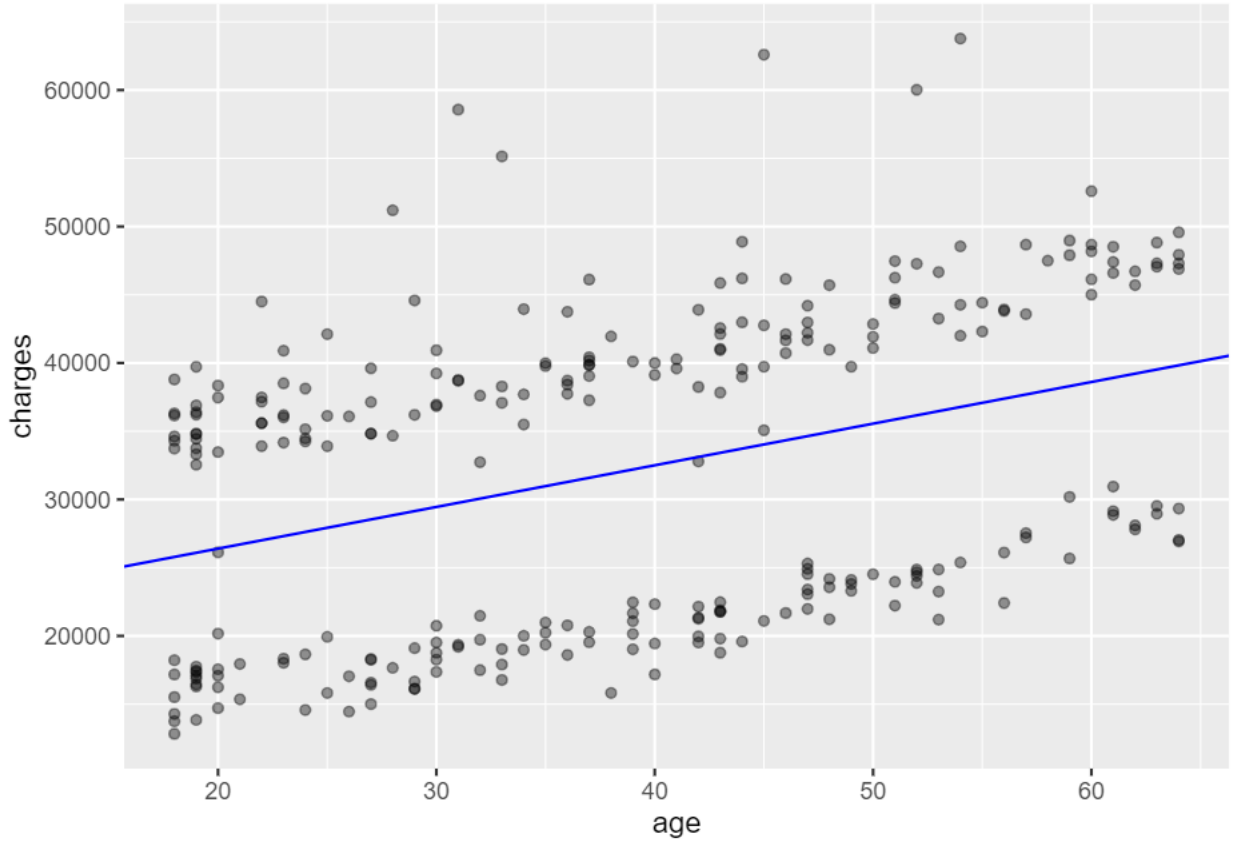
Next, we create the charges v age model and assess the assumptions

```
# Follow the same steps for model creation and assumption checking

age_model = lm(charges ~ age, data = insurance_smokers)
summary(age_model)
```

```
##
## Call:
## lm(formula = charges ~ age, data = insurance_smokers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -16072 -11137   5764   8592  28815
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20294.13    1913.40  10.606  < 2e-16 ***
## age           305.24      46.73   6.532 3.18e-10 ***
## ---
```

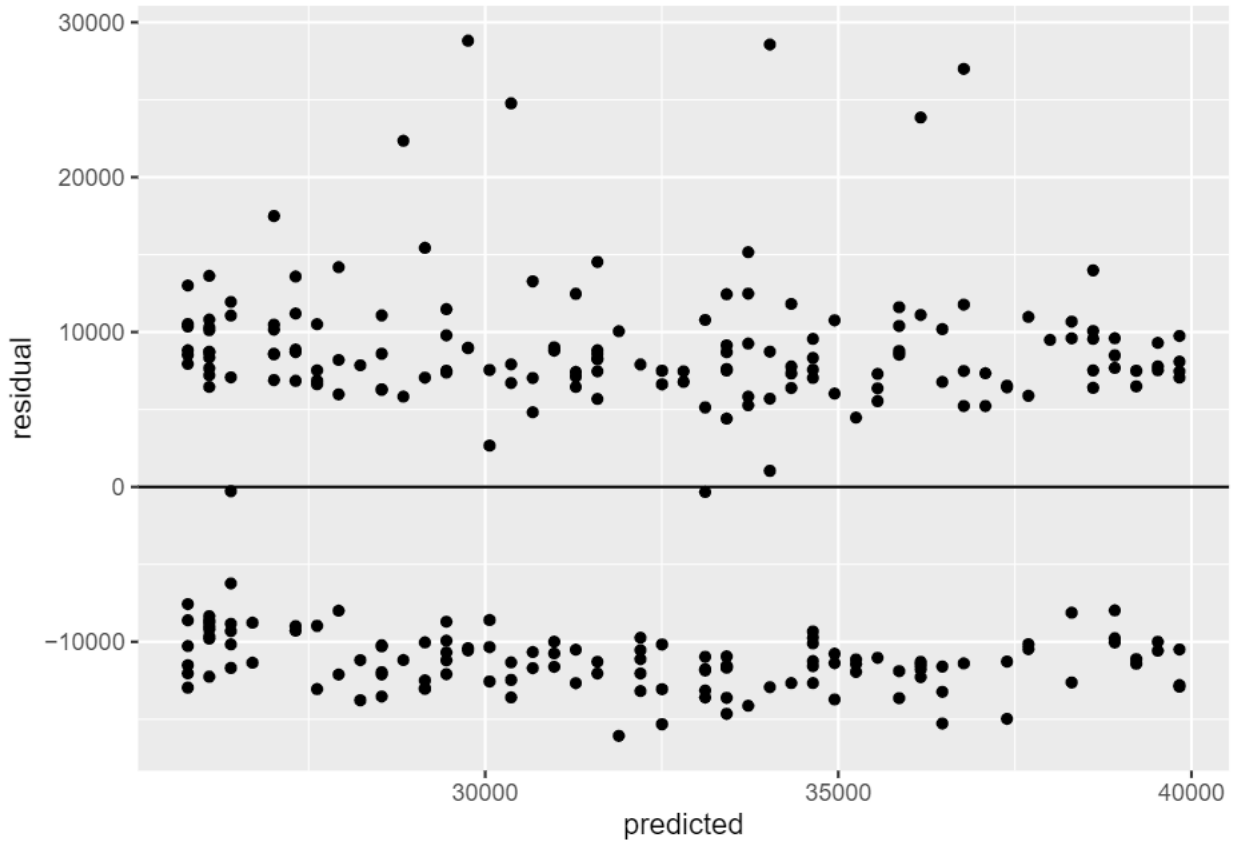
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10750 on 272 degrees of freedom
## Multiple R-squared:  0.1356, Adjusted R-squared:  0.1324
## F-statistic: 42.67 on 1 and 272 DF,  p-value: 3.181e-10
```

```
# Assumption 1 is met because it is reasonable to assume one person's age and
# insurance charge is independent from another's

ggplot(insurance_smokers, aes(y = charges, x = age)) +
  geom_point(alpha = 0.4) +
  geom_abline(slope = 305.24, intercept = 20294, color = "blue")
```

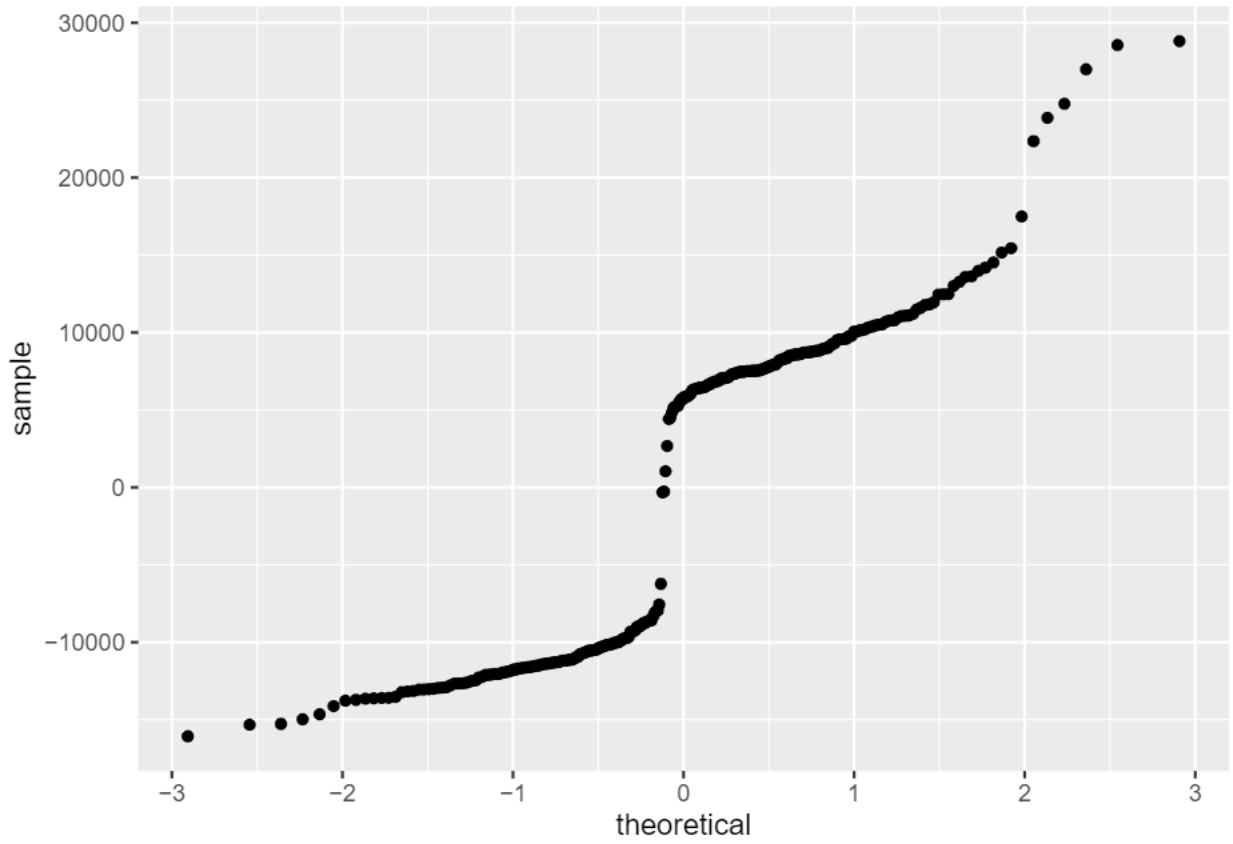



```
# It is interesting to note that there are two distinct groups in the data,
# each of which seems to have a nearly identical linear relationship.
# None of the other variables in the dataset can account for these two groups,
# so their cause is unknown, but it most likely has something to do with risk.
# We sould expect the residual plot to be symmetric and randomly distriuted,
# but focused towards values of y farther from 0, since the model produces an
# estimate that is in the middle of the two distinct groups.

age_mod_results = data.frame(observed = insurance_smokers$charges,
                             predicted = age_model$fitted.values,
                             residual = age_model$residuals)

# Plot the residuals versus the predictions
ggplot(age_mod_results, aes(y = residual, x = predicted)) +
  geom_point() +
```

```
  geom_hline(yintercept = 0)
```




```
# The residual plot looks exactly as expected and described in the previous
# block comment, with residual values far from 0, but otherwise randomly
# and symmetrically distributed


# Create a Q-Q plot of the residuals
ggplot(age_mod_results, aes(sample = residual)) +
  geom_qq()
```




```
# The Q-Q plot consists of 2 distinct sections, one for negative residual
# values and one for positive, each of which looks like a moderate straight
# line (and therefore fulfills assuption of 2). To make the residual and
# Q-Q plots perfect, one must subset the data by the two distinct groups in the
# scatterplot, which cannot be done given that there is no obvious reason
# for these 2 groups in the dataset. Because of the issues in the residual
# and Q-Q plots, neither of these models perfectly fulfill assumption 2.
```

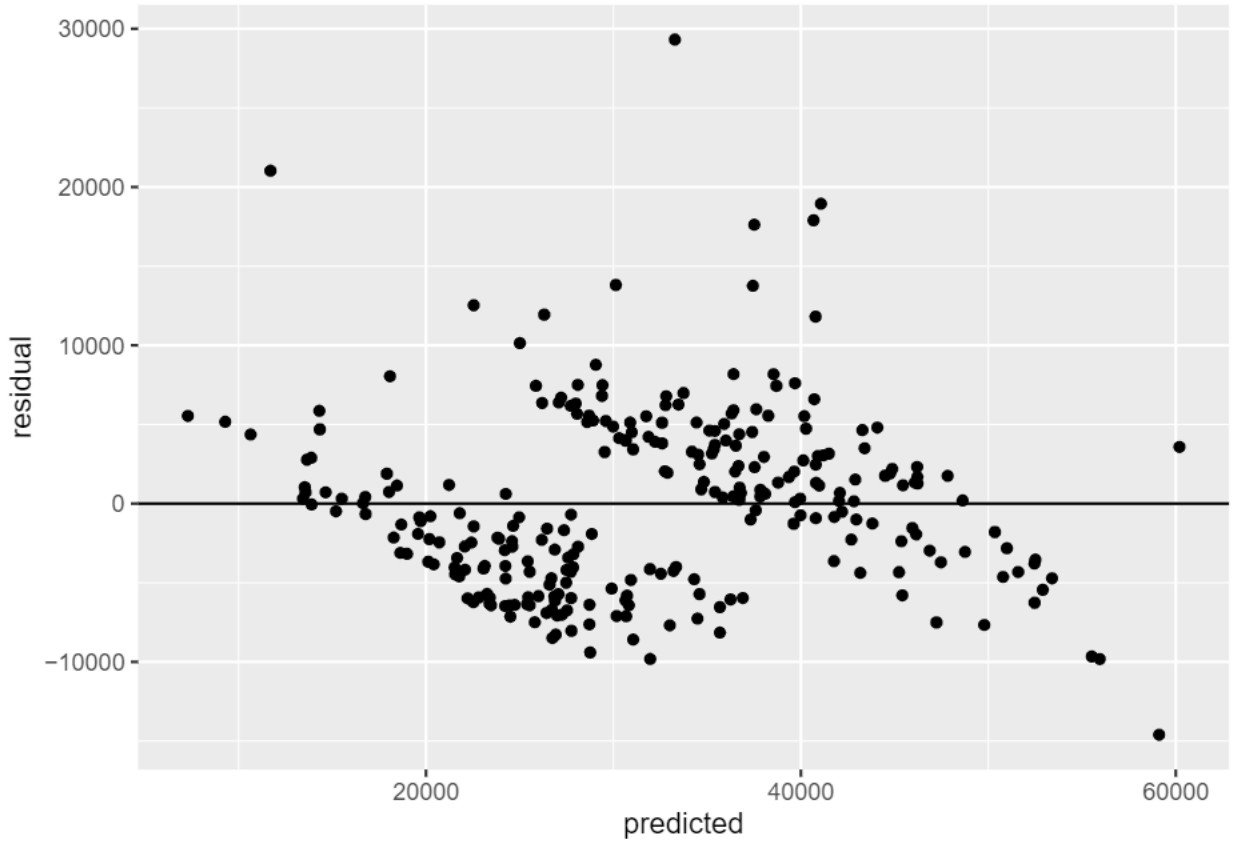
Next, we create a multivariate linear model using both age and bmi

```
# Follow the same steps for model creation and assumption checking

combined_model = lm(charges ~ age + bmi, data = insurance_smokers)
summary(combined_model)
```

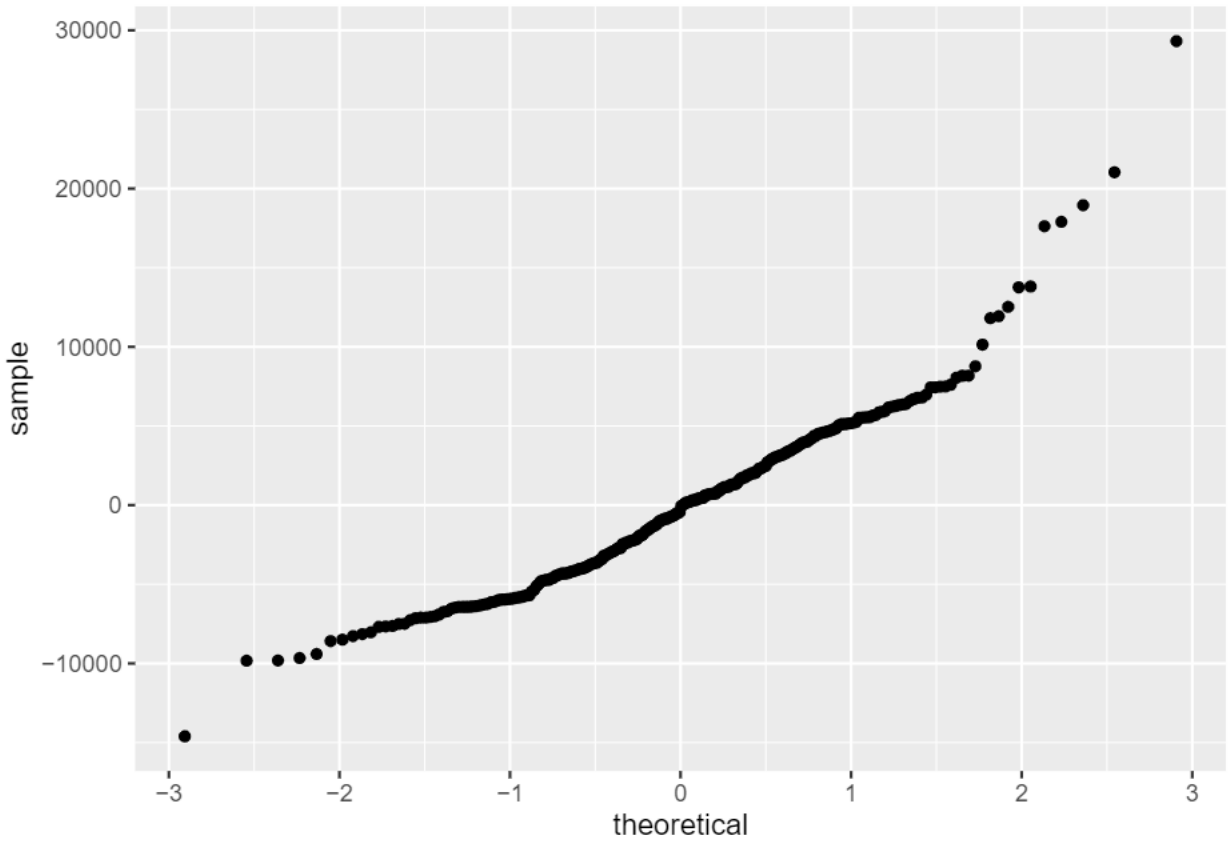
```
##
## Call:
## lm(formula = charges ~ age + bmi, data = insurance_smokers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14604.4  -4315.1   -240.5   3638.0  29316.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22367.45    1931.86  -11.58   <2e-16 ***
## age            266.29      25.06   10.63   <2e-16 ***
## bmi           1438.09      55.22   26.05   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5754 on 271 degrees of freedom
## Multiple R-squared:  0.7532, Adjusted R-squared:  0.7514
## F-statistic: 413.6 on 2 and 271 DF,  p-value: < 2.2e-16
```

```r
# Assumption 1 is met because it is reasonable to assume age and bmi and
# insurance charge of a person is independent from another's.

combined_mod_results = data.frame(observed = insurance_smokers$charges,
                                  predicted = combined_model$fitted.values,
                                  residual = combined_model$residuals)

ggplot(combined_mod_results, aes(y = residual, x = predicted)) +
  geom_point() +
  geom_hline(yintercept = 0)
```




```r
# The residual plot has values symmetrically distributed about 0, with low y
# values, however there is in obvious pattern, with a break in the middle
# most likely caused by the break in the age. However, of all the residual
# plots, this fits the requirements the closest.


# Create a Q-Q plot of the residuals
ggplot(combined_mod_results, aes(sample = residual)) +
  geom_qq()
```




```
# The Q-Q plot for the model of combined bmi and age looks correct, with
# a single straight line, signifying that the combined model meets assumption
# 2 better than any of the other previous models
```

To conclude our work with these models, we assess the quality of each model created using MSPE and R-squared.The R-squared values are taken from the summaries printed above.

```
# bmi model

#MSPE
mean(bmi_mod_results$residual^2)
```

```
## [1] 46397859
```

```
"R^2: 0.6504"
```

```
## [1] "R^2: 0.6504"
```

```
# lowbmi model

# MSPE
mean(lowbmi_mod_results$residual^2)
```

```
## [1] 22935026
```

```
"R^2: 0.09283"
```

```
## [1] "R^2: 0.09283"
```

```
# highbmi model

# MSPE
```

```
mean(highbmi_mod_results$residual^2)
```

```
## [1] 29246158
```

```
"R^2: 0.1333"
```

```
## [1] "R^2: 0.1333"
```

```
# age model

# MSPE
mean(age_mod_results$residual^2)
```

```
## [1] 114725593
```

```
"R^2: 0.1356"
```

```
## [1] "R^2: 0.1356"
```

```
# combined model

# MSPE
mean(combined_mod_results$residual^2)
```

```
## [1] 32750237
```

```
"R^2: 0.7532"
```

```
## [1] "R^2: 0.7532"
```

```
# From this information, especially that of the R^2 values, combined
# with information from the assumptions section, we conclude that the
# combined model is by far the best model for linear prediction because it has
# the highest R^2 value, and fits the assumptions the closest (although not
# perfectly)
```

To conclude the project, we will use a 90% confidence interval to predict the average cost of insurance for a nonsmoker older than 40 with a bmi under 30.

```
# Create a dataset of nonsmokers older than 40 with a bmi under 30

confidence_data = insurance %>%
  filter(smoker == "no") %>%
  filter(bmi < 30) %>%
  filter(age >= 40)

dim(confidence_data)
```

```
## [1] 226   8
```

```
# Set a random number seed

set.seed(2021)

# Take a sample of 150 people from the dataset

sample = sample(confidence_data, size = 150, replace = FALSE)

head(sample)
```

```
##      age    sex    bmi children smoker    region   charges logcharges orig.id
```

```
## 135  60   male 24.320        0     no northwest 12523.605   9.435371     135
## 166  48   male 29.600        0     no southwest 21232.182   9.963273     166
## 174  49   male 29.830        1     no northeast  9288.027   9.136481     174
## 186  49 female 29.925        0     no northwest  8988.159   9.103663     186
## 140  51   male 25.400        0     no southwest  8782.469   9.080513     140
## 70   51 female 20.600        0     no southwest  9264.797   9.133977      70
```

```
# Calculate the sample mean charge

sample_mean = mean(sample$charges)
sample_sd = sd(sample$charges)

print(sample_mean)
```

```
## [1] 11436.41
```

```
print(sample_sd)
```

```
## [1] 5096.105
```

```
# n > 40, so we are able to calculate the confidence interval for the mean
# charge using the sample mean and standard deviation, using the z value for
# a 90% CI, 1.645, and the formula:
# x_bar ±1.645(s/sqrt(n))

lower_bound = sample_mean - 1.645 * (sample_sd/sqrt(150))
upper_bound = sample_mean + 1.645 * (sample_sd/sqrt(150))

print(lower_bound)
```

```
## [1] 10751.93
```

```
print(upper_bound)
```

```
## [1] 12120.89
```

```
# The 90% confidence interval for expected charge for a nonsmoker with bmi
# under 30 and age over 40 is [10751.93, 12120.89]
```

Conclusion:

The results, especially that of the R^2 values, and the information from the assumptions demonstrate that the best possible model is the combined model and it is by far the best model for linear prediction because it has the highest R^2 value, and fits the assumptions the closest (although not perfectly). However, the charges v age model did provide some interesting results as well.

In the charges v age model we saw that there are two distinct groups in the data, each of which seems to have a nearly identical linear relationship. However, none of the other variables in the dataset can account for these two groups, so their cause is unknown. This model, however, did have the best residual plot and if the separating factor among the two groups were to be researched in depth, it seems that the Q-Q plot for the separated models would also be promising.

Although, age is a pretty primitive and natural factor for charges so the usefulness of this model would be relatively lower than the others, but still helpful.