Homework_5

Bailey Ho | Daria Barbour-Brown | Warren Kennedy

2023-05-03

Conceptual Problem

Question 1

Why do we need to check modeling assumptions for linear regression?

We need to check the modeling assumptions of linear regression in order to ensure the validity of statistical inference, identify potential issues with the data, and guide model selection and improvement. The key model assumptions for linear regression are linearity, homoscedasticity, uncorrelated errors, no significant outlines, and normally distributed errors. If the data is not linear, it does not make sense to use linear regression to describe relationships in the data. If the errors correlated, heteroskedastic, or do not come from a normal distribution, then our estimates may be biased and inefficient, leading to inaccurate conclusions about the relationships between the predictor and the response. Finally, if our data contains significant outliers, then the linear estimation will likely be subject to inaccurate influence due to the inflexibility of plotting a linear relationship that fits all of the data.

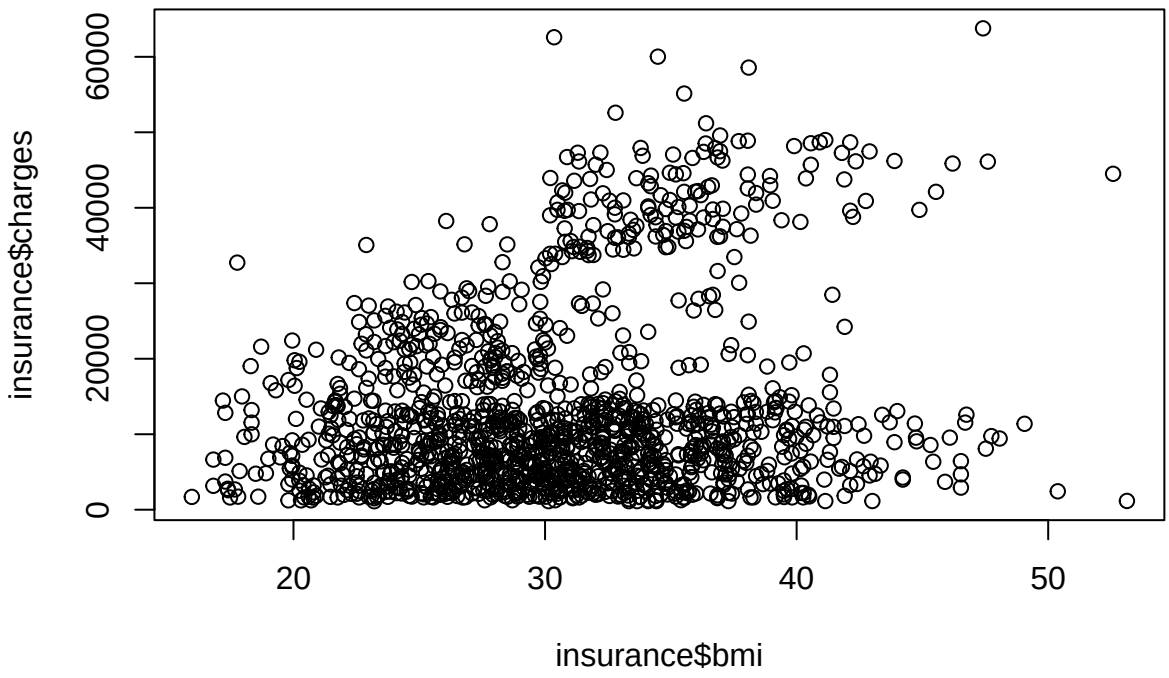
Furthermore, if the assumptions are met, a linear regression model will be the best linear unbiased estimate for our data which allows the data analyst to make strong statistical inferences.

Aplication Questions

Question 2

Much of the data seems to be restricted to lower values of the dependent variable. Despite this, there also seems to be a positive linear relationship between charges and bmi albeit with a smaller subset of the data.

```
insurance <- read.csv("homework4_insurance.csv")
plot(insurance$bmi, insurance$charges)
```

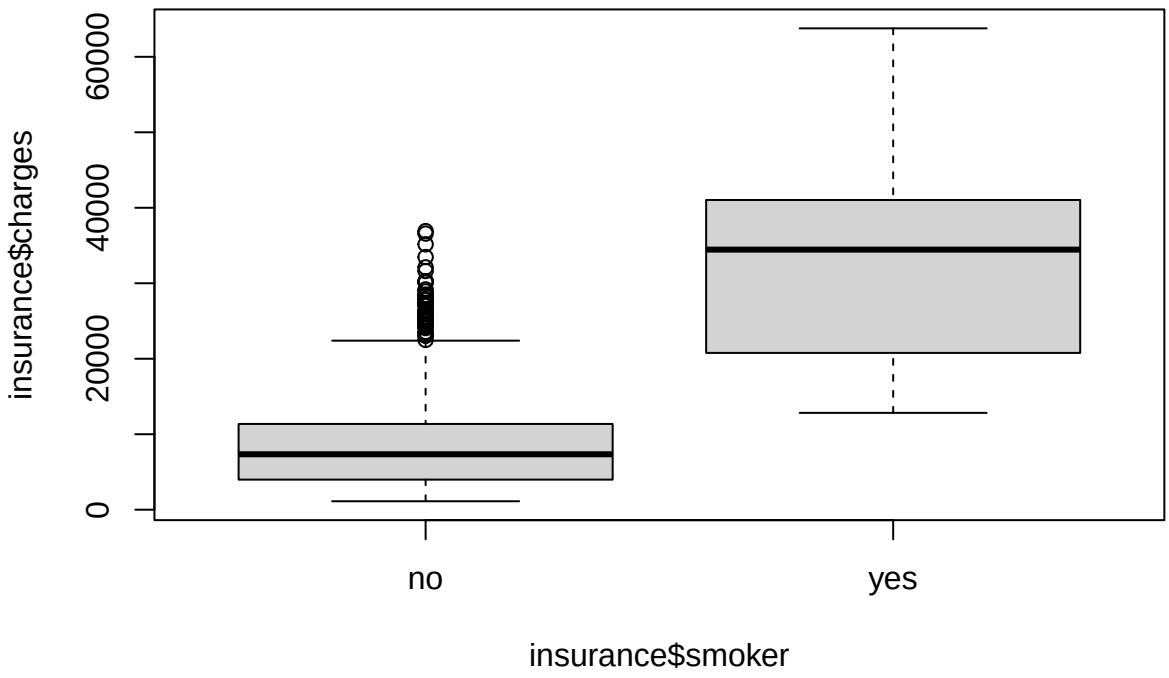


Question 3

The median value for insurance charges to non-smokers is significantly below the median charges for smokers. With this we conclude that smokers have higher medical bills than non-smokers.

```
smokers <- insurance$smoker[insurance$smoker=='yes']
non_smoker <- insurance$smoker[insurance$smoker=='no']

boxplot(insurance$charges ~ insurance$smoker, height = .6)
```



Question 4

(a)

```
linearFitBmi_Smoker <- lm(charges~bmi+smoker, data = insurance)
summary(linearFitBmi_Smoker)
```

```
##
## Call:
## lm(formula = charges ~ bmi + smoker, data = insurance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15992.7  -4600.2   -802.4   3636.2  30677.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3459.10     998.28  -3.465 0.000547 ***
## bmi           388.02      31.79  12.207  < 2e-16 ***
## smokeryes   23593.98     480.18  49.136  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7088 on 1335 degrees of freedom
## Multiple R-squared:  0.6579, Adjusted R-squared:  0.6574
## F-statistic:  1284 on 2 and 1335 DF,  p-value: < 2.2e-16
```

(b)

The summary returns 23593.98 as the coefficient for smoker. If you are a smoker you will be charged $23593.98 more that non-smokers.

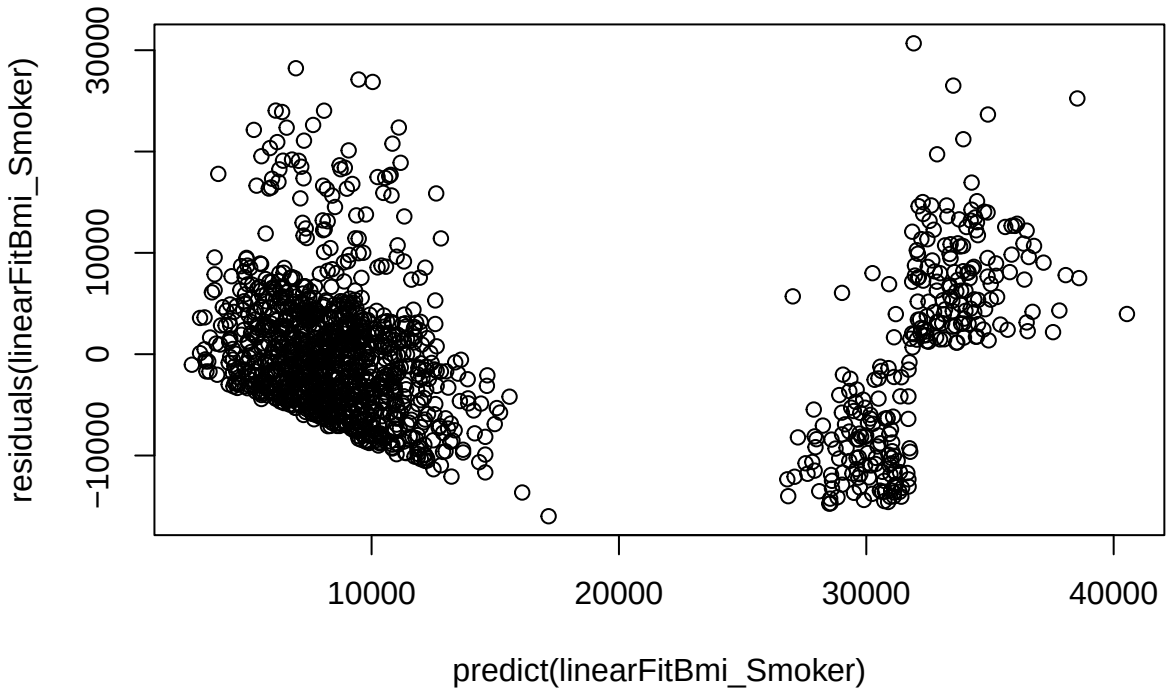
(c)

The summary returns a $388.02 coefficient for bmi. This coefficient explains the marginal change in medical charges for each unit increase/decrease of bmi.

(d)

We observe that the residuals are not centered around zero and are instead clustered around various predicted values. This could mean the errors are correlated and that there may be a pattern in the data that our model is not picking up. This may violate the assumption of homoscdasticity, normally distributed errors, and/or uncorrelated errors. The data may also possibly violate our linearity assumption.

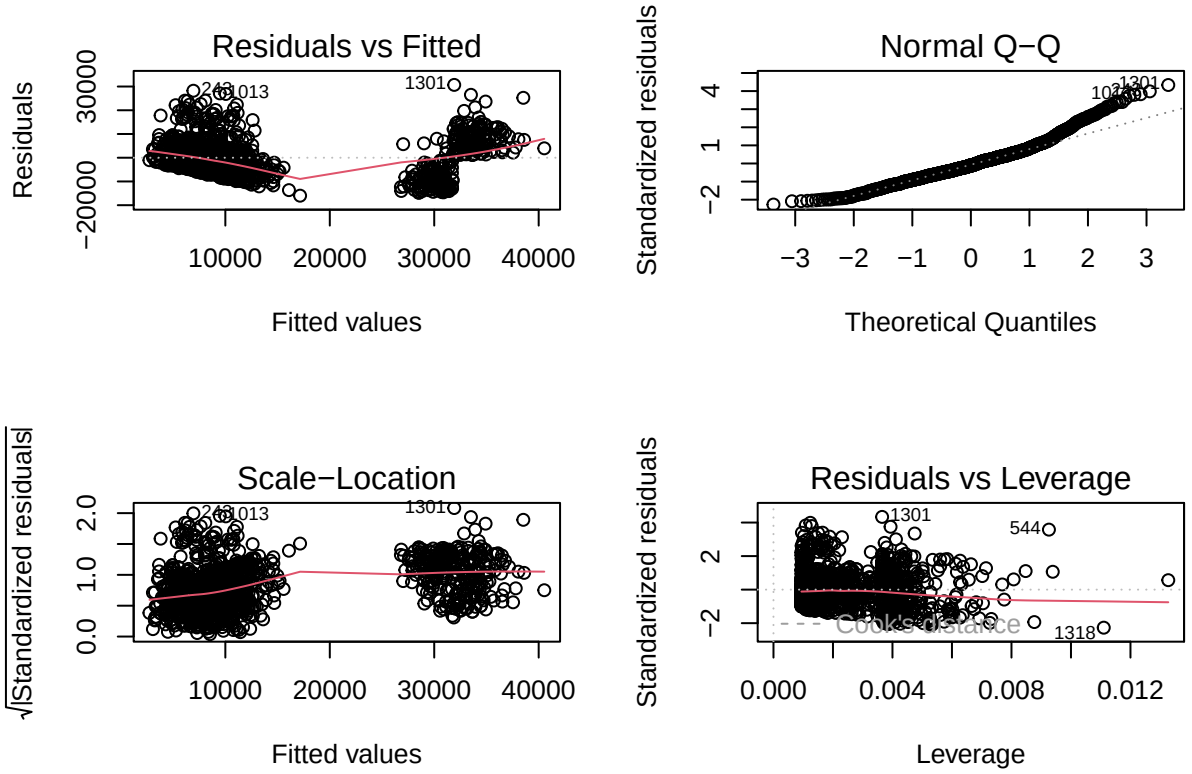
```
plot(predict(linearFitBmi_Smoker), residuals(linearFitBmi_Smoker))
```



(e)

The Normal Q-Q plot is a visual plot ofresidual normal distribution. If the points fall along the dasched line within the plot, the the residuals are approximately normally distributed. If the points deviate from a straight line, the data may not be normally distributed. We see that, for both low and high values of theoretical quantiles, the residuals deviate from the dashed line suggesting that our residuals also violate the assumption of normality.

```
par(mfrow = c(2, 2))
plot(linearFitBmi_Smoker)
```



Question 5

(a)

```
logLinearFitBmi_Smoker <- lm(log(charges)~bmi+smoker, data = insurance)
summary(logLinearFitBmi_Smoker)
```

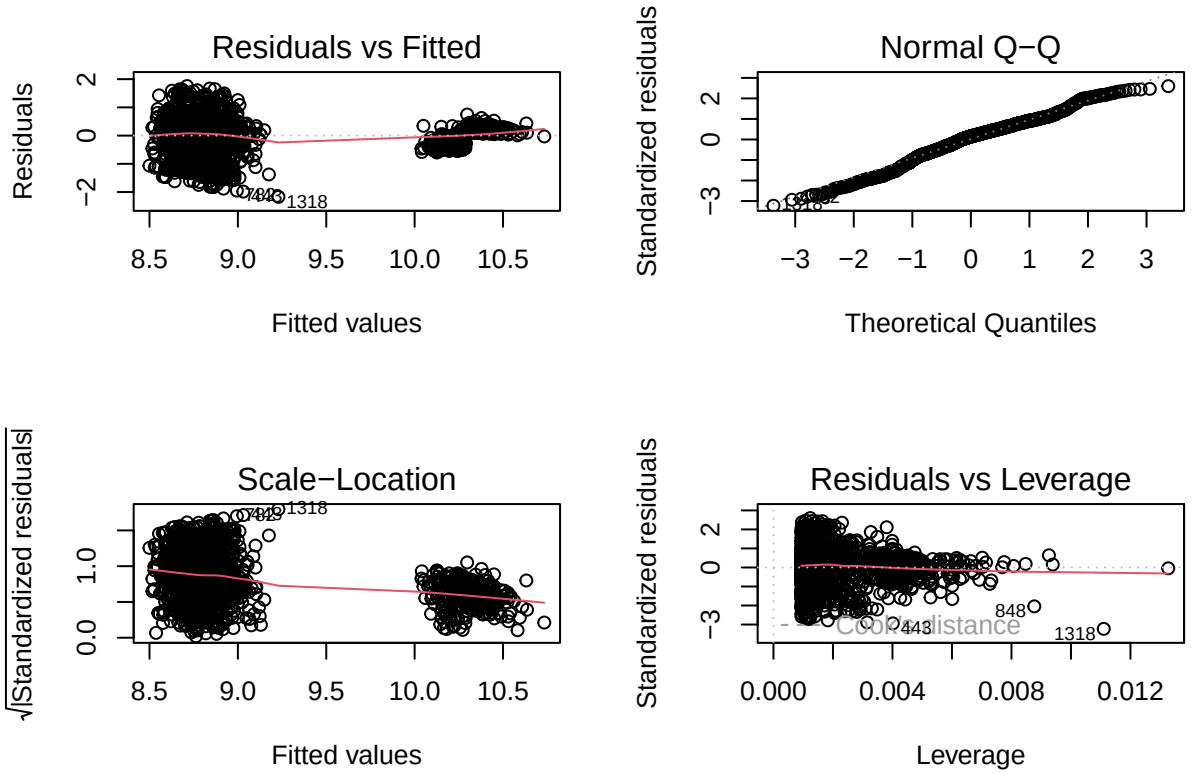
```
##
## Call:
## lm(formula = log(charges) ~ bmi + smoker, data = insurance)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -2.17030 -0.40754  0.09871  0.44368  1.75504
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.186576   0.095254  85.945  < 2e-16 ***
## bmi         0.019629   0.003033   6.472 1.36e-10 ***
## smokeryes   1.514765   0.045818  33.061  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6763 on 1335 degrees of freedom
## Multiple R-squared:  0.4598, Adjusted R-squared:  0.459
## F-statistic: 568.3 on 2 and 1335 DF,  p-value: < 2.2e-16
```

(b)

1.514765 is the marginal percent change of medical charges for smokers. This means you will be charged 150% more as a smoker than as a nonsmoker. ## (c) In this case, the assumption of normality holds.

```
par(mfrow = c(2, 2))
plot(logLinearFitBmi_Smoker)
```



(d)

The residual plots from the model in question 4 are nonlinear in both high and low theoretical quantiles suggesting the residuals do not come from a normal distribution. After taking the log of the dependent variable, our updated model shows that all of the residuals are now plotted close to the dashed line. This updated plot suggests that the our data has normally distributed residuals.

Question 6

(a)

Suppose we want to test whether or not BMI has a significant effect predicted medical charges. Our **null hypothesis** would be that BMI has no effect on predicted medical charges. Our **alternative hypothesis** would be that BMI has a statistically significant effect on predicted medical charges. Our **dependent** variable would be medical charges. The **independent variable** would be BMI.

(b)

Testing our hypothesis at the 95 percent significance level, we conclude that BMI has a statistically significant effect on medical charges due to the p-value being lower that our alpha.

```
Htest1= lm(charges~bmi,data=insurance)
summary(Htest1)
```

```
##
## Call:
## lm(formula = charges ~ bmi, data = insurance)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -20956  -8118  -3757   4722  49442
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1192.94    1664.80   0.717    0.474
## bmi           393.87      53.25   7.397 2.46e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11870 on 1336 degrees of freedom
## Multiple R-squared:  0.03934,    Adjusted R-squared:  0.03862
## F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

(c)

```
Htest1CI <- confint(Htest1,level=0.95)
Htest1CI
```

```
##                   2.5 %    97.5 %
## (Intercept) -2072.9743 4458.8487
## bmi           289.4089  498.3372
```