Lab 11: Checking assumptions for linear regression

#Part I:

Boston data on median household value and air pollution

We are examing a data set used to predict housing prices in the area around Boston (Harrison, D. and Rubinfeld, 1978). We wish to examine specifically the association of the measure of housing price (`medv`, median value of owner-occupied homes in the $1000s) and a measure of air pollution (`nox`, nitrogen oxides concentration, parts per 10 miillion). The data frame (Boston) is contained in another package (MASS), which we load below.

```
library(broom)
library(dplyr)
library(ggplot2)
library(tidyr)

# load library with data
library(MASS) #note: this package has a function `select()` that overrides
# dplyr's select function that we use below. To ensure we use the dplyr version,
# we specify `dplyr::select()` when we use the function

# list variables
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

```
# variable definition - take a quick look at the variables in the data frame
#help(Boston)

# Examine data
head(Boston) #or View(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

1. Perform and summarize the results of a linear regression of `medv` versus `nox` using Boston data. Be careful about which variable is explanatory and which is response!

```
# write your code here.
```

```
lm_Boston <- lm(medv ~ nox, data = Boston)
tidy(lm_Boston)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)     41.3      1.81      22.8 9.87e-80
## 2 nox            -33.9      3.20     -10.6 7.07e-24
```

2. Recall that before we only interpreted the estimate for the intercept and slope parameter. Interpret the slope parameter for `nox`. Notice also that the other columns are `std.error`, `statistic`, and `p-value` – these should remind you things we've learned during Part III of the course on inference. They correspond to the hypothesis test with the null hypothesis that the parameter is equal to 0. Thus, how would you interpret the p-value for `nox`? (We will cover this in detail in class on Friday.)

<Write your answer here.>

3. Use `glance()` to look at the r^2 value for this model. Does `nox` explain alot of the variance in median household value? Would you expect it to?

```
#Put your code here
glance(lm_Boston)
```

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl> <dbl> <dbl>
## 1     0.183         0.181  8.32      113. 7.07e-24     2 -1789. 3584. 3597.
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

<Write your answer here.>

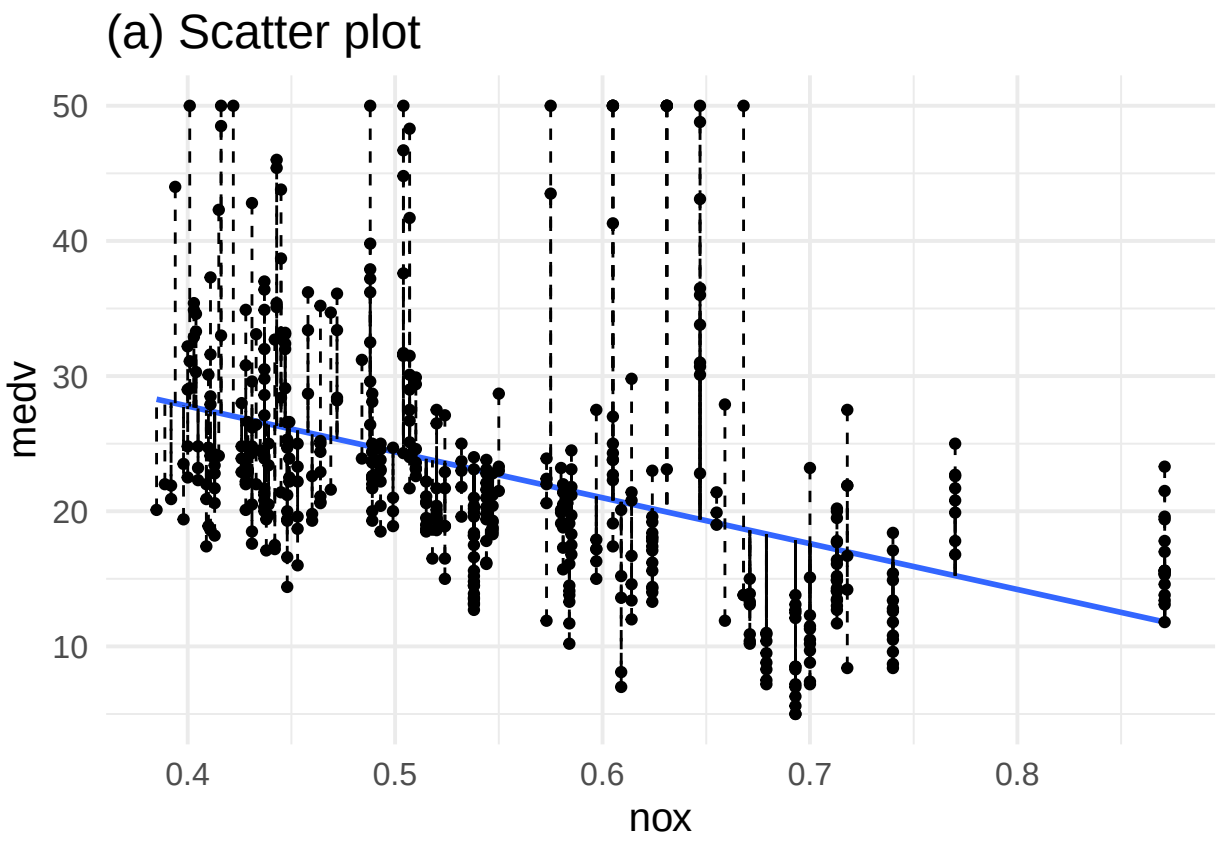
`nox` explains 18.3% of the variation in median household value. Thus it does capture some of the variation in value. We wouldn't expect it to capture too much, because there are many other factors that also explain household value such as characteristics of the neighborhood, such as the income levels of people living in the neighborhood, or the crime rates occurring in a neighborhood. Note that these characteristics are all related/associated with household value, and not necessarily causes of household value.

4. Check the assumptions required for the simple linear model using the plots shown during Wednesday's lecture. Note that to make these plots you first need to fit the linear model and then use the `augment()` function from the broom package to store the residuals and fitted values into a new data frame. Try to update the plotting code from class for your data set. The GSIs can help you if you get stuck!

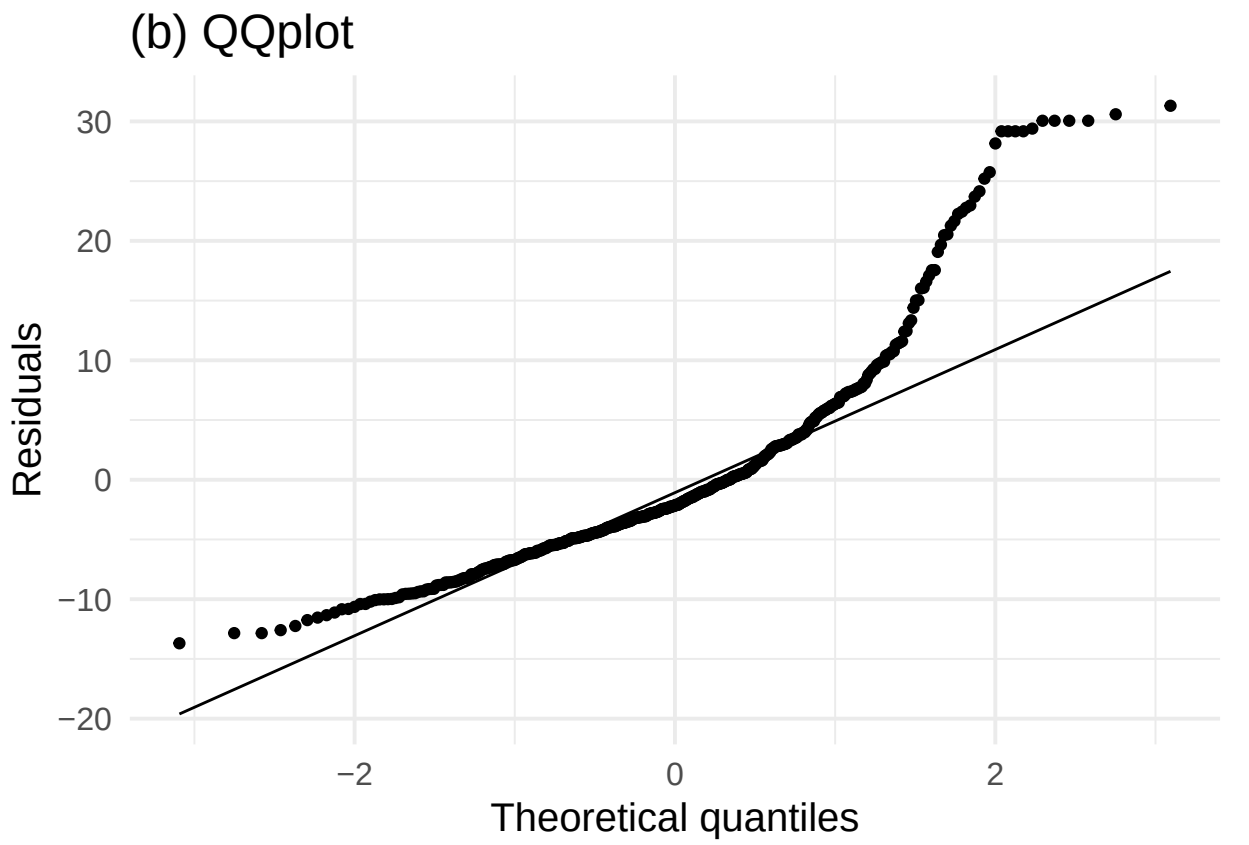
The hardest one to make is likely the boxplots because the data first needs to be reshaped to make the plots. The reshaping code is the `gather()` function that you see in the slides/Rmd file for making these plots. Here is a helpful R explainer for how gather() works: https://twitter.com/WeAreRLadies/status/1059520693857996800. Basically, we need to gather the observed y values and the residuals by stacking them into one variable so that we can make two box plots side-by-side. Below, we include the `gather()` code for you since it is a bit tricky. You need to use the resulting data frame to make your box plots.

```
# Put your code here

# augment your model

# first plot

# second plot

# third plot
```

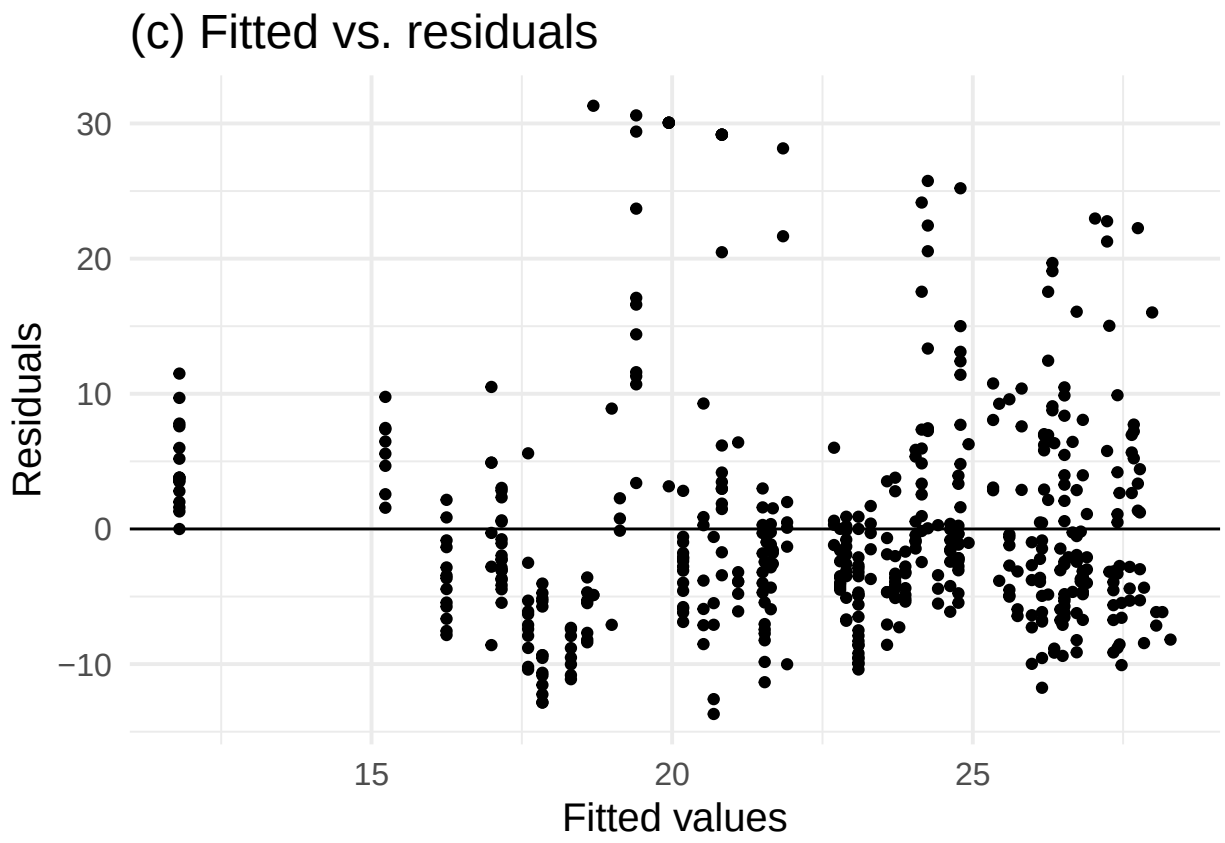
```r
# fourth plot (gather could included for you. It assumes your augmented data is
# called `augmented_1`, so you will likely need to update that to whatever your
# augmented data is called.)
#reshape <- augmented_1 %>% dplyr::select(.resid, medv) %>%
#  gather(key = "type", value = "value", medv, .resid)

#use patchwork to combine the plots into one figure.

# solutions
augment_Boston <- augment(lm_Boston)

## Fitted model
plot1 <- ggplot(augment_Boston, aes(nox, medv)) +
  geom_smooth(method = "lm", se = F) +
  geom_point() +
  geom_segment(aes(xend = nox, yend = .fitted), lty = 2) +
  theme_minimal(base_size = 15) +
  labs(title = "(a) Scatter plot")

# QQ plot
plot2 <- ggplot(augment_Boston, aes(sample = .resid)) +
  geom_qq() +
  geom_qq_line() +
  theme_minimal(base_size = 15) +
  labs(y = "Residuals", x = "Theoretical quantiles", title = "(b) QQplot")

## Fitted vs. residuals
plot3 <-ggplot(augment_Boston, aes(y = .resid, x = .fitted)) +
  geom_point() +
  theme_minimal(base_size = 15) +
  geom_hline(aes(yintercept = 0)) +
  labs(y = "Residuals", x = "Fitted values", title = "(c) Fitted vs. residuals")

## Amount explained
reshape <- augment_Boston %>% dplyr::select(.resid, medv) %>%
  gather(key = "type", value = "value", medv, .resid)

plot4 <- ggplot(reshape, aes(y = value)) +
  geom_boxplot(aes(fill = type)) +
  labs(title = "(d) Amount explained") +
  theme_minimal(base_size = 15)

plot1
```



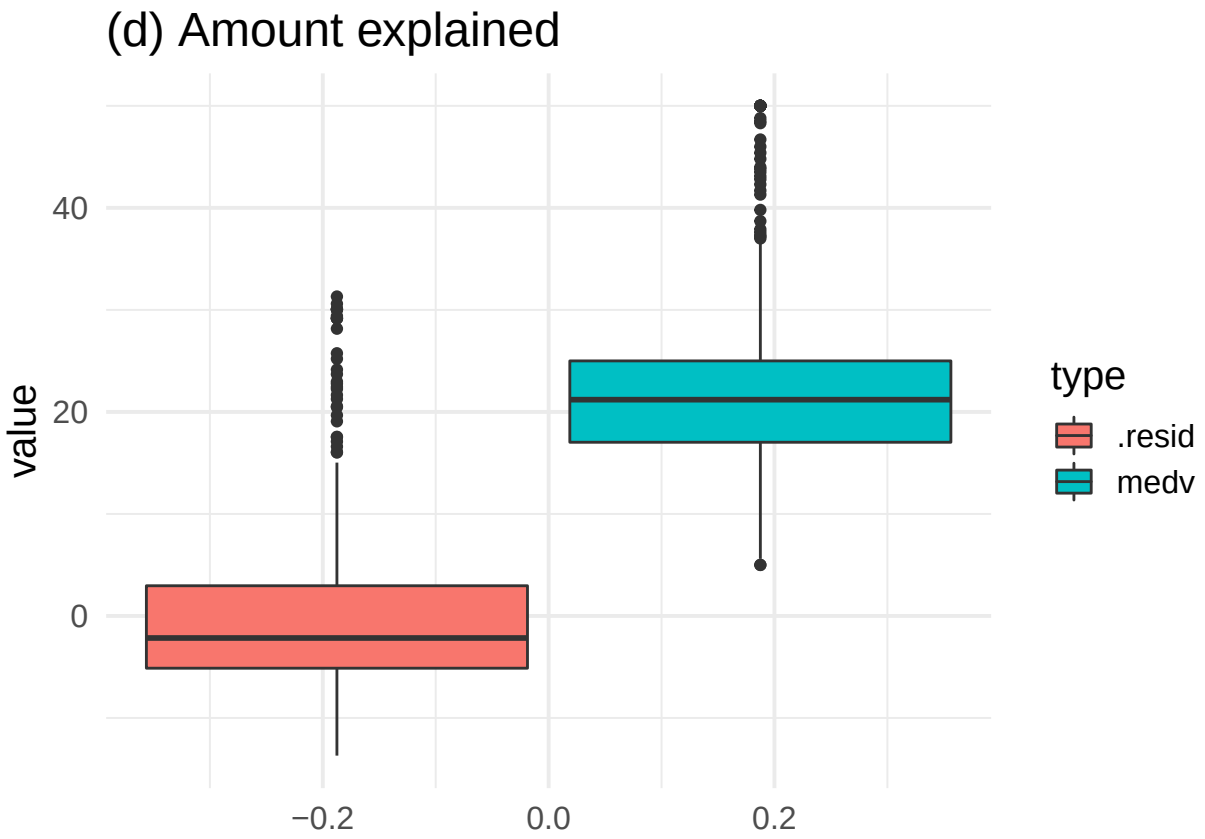

```
plot2
```


(b) QQplot
Residuals
30
20
10
0
−10
−20
−2
0
2
Theoretical quantiles

```
plot3
```


(c) Fitted vs. residuals
Residuals
30
20
10
0
−10
15
20
25
Fitted values

```
plot4
```



5. What do you think about the assumption plots? They are a bit messier than those shown in class, and reflect what we often see "in the real world".

<Write your answer here.>

- (a): There are many points that appear to be much higher than the curve, or away from the many "cloud" of the data. It also looks like there is an upper bound at `medv=50`. As Peter mentioned in class, this could indicate that there is some upper bound in the data frame for some reason, something we didn't notice without plotting the data! There does appear to be a negative association between nitrogen oxide and median household value, where higher concentrations are associated with lower median values.
- (b): The QQ plot is capturing non-Normality of the residuals. The shape of the QQ plot is showing the large positive residuals corresponding to the high median values that aren't predicted by the model.
- (c): The fitted vs. residuals is not a "random cloud" like we would prefer. It is tricky to interpret – you might see a bit of a curve pattern because reading left to right the residuals start all above 0, then drop below 0 (generally) and then go back up. But, there are all those positive outliers with very large residuals. Seeing the modest quadratic shape may convince you to try adding a quadratic term to the model and see if that model performs better.
- (d): Yikes. The IQR of the residuals is basically the same size as the IQR of the y variable itself. The model does not do a great job explaining the variation in median value. This makes sense because the only variable in our model is Nitrogen oxide – think of all the other things that could predict household value. In multiple linear regression, the model is expanded to include other variables, but we don't go into this in PH142.

Lab Conclusion (make sure to read this and understand it)**

From this exercise, we can conclude that there is a negative association between air pollution and median household value. An increase in Nitrogen Oxide of 1 parts per million (PPM) is associated with a decrease in median household value of $33,900 (see `help(Boston)` to remind yourself of the units for `nox` and `medv`). Note that this "increase of 1 unit" is wider the range of the scatter plot, so we should modify it to talk about an increase of 0.1 unit in `nox`. That is, an increase in Nitrogen Oxide of 0.1 PPM is associated with a decrease in median household value of $3,390. This you can see when you look at the scatter plot of the data and the line of best fit. – Look at going from 0.5 to 0.6 on the x axis and see how the model predicted y going from ~$25k to ~$22k.

#Part II

1. Perform and summarize the results of a linear regression of `medv` (median value of owner-occupied homes in $1000s) and `dis` (weighted mean of distances to five Boston employment centres) using Boston data. Be careful about which variable is explanatory and which is response!

```
dat = Boston
# write your code here.

# solution
lm_dis <- lm(medv ~ dis, data = dat)
tidy(lm_dis)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    18.4      0.817     22.5  4.01e-78
## 2 dis             1.09     0.188      5.79 1.21e- 8
```

2. Interpret the slope parameter and p-value from the table. What null and alternative hypotheses does this p-value refer to?

Solution H_0: slope parameter is equal to 0 H_a: slope parameter is not equal to 0

For every 1 unit increase in the distance to an employment center (units unspecified), the median value of owner-occupied homes increases by $1,190.

A p-value of 0.0262 implis a 2.6% chance of observing this slope coefficient under the null hypothesis. This provides evidence in favour of the alternative hypothesis that there is an association between distance to an employment center and median household value.

3. Derive a 95% CI for this slope parameter. In your opinion, would you expect the direction of this relationship to hold if the data were collected today?

95%CI: slope estimate +/- t*SE

```
t_star <- qt(p = 0.975, df = 48)
```

95%CI: 1.19 +/- 2.01*0.520 = 0.1450209 to 2.234979

4. Use a function to look at the r-squared value for this model. Does `dis` explain alot of the variance in median household value? Would you expect it to?

```
#Put your code here

#Solution
glance(lm_dis)
```

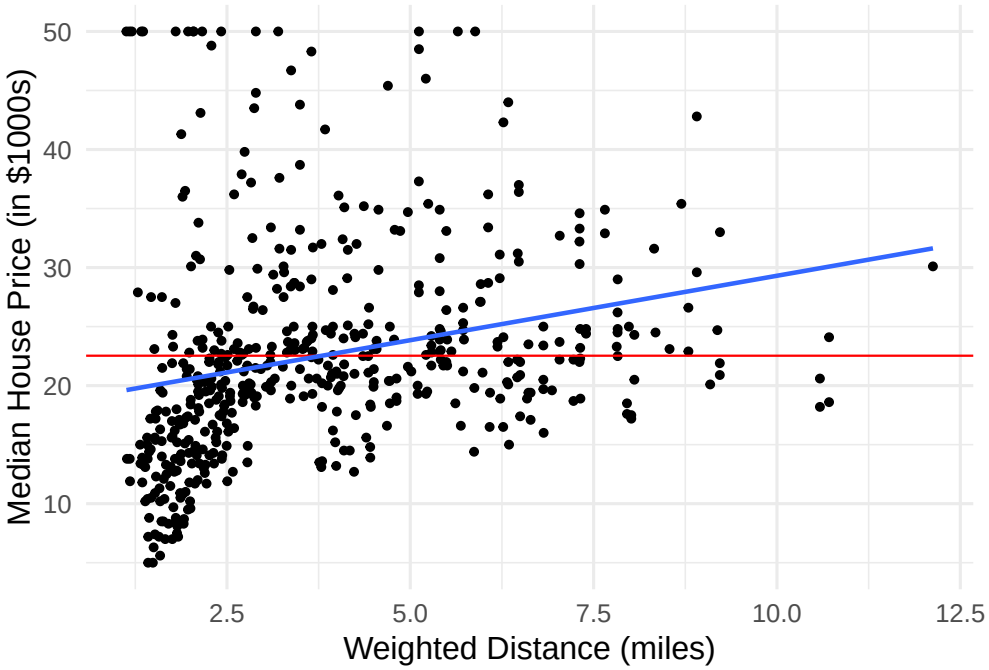
```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik   AIC   BIC
```

```
##       <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl> <dbl> <dbl>
## 1    0.0625        0.0606  8.91      33.6 1.21e-8     2 -1824. 3654. 3667.
## # ... with 2 more variables: deviance <dbl>, df.residual <int>
```

<Write your answer here.>

`dis` explains 9.8% of the variation in median household value. Thus it does capture some, but very little of the variation in value. However, it is only one of many factors that predict housing prices so we not expect the r-squared to be very high. Also, note that these characteristics are all related/associated with household value, and not necessarily causes of household value.

5. Back to the fit of the model of `medv` vs. `dis`. Make a plot with the raw data points, the fitted line from the simple linear regression model (only containing `medv` and `dis`) and also add a line with a slope of 0. You can have that line cross the y axis at the average value of `medv` to vertically bisect the data points.



6. For you, does the plot raise any concerns about the assumptions of the linear regression you just performed? What other plots might you do to explore the fit?

<Write your answer here.>

Solution: - Might say that there is less variation in the residuals for longer distances. But this could also be because there is little data > 6 miles. The linear relationship is not very strong in any case, which is why the r-squared for the model is so low. Despite this, we can still see some association because as distance increases the cloud of points for median price also increases. - students might also note that when distance < 2.5 the residuals are mostly negative, suggesting a violation of the assumption of Normally-distributed residuals for each value of x. This might indicate that a curved model would fit better.

Regardless of your answer, we go forward using the model to make inferences about the points on the line.

###(Optional) Pointwise Confidence Intervals and Multiple Testing

As you learned in lecture, there are two types of confidence intervals applicable to estimating a point on the plot which are related to whether one is predicting the population average among individuals with $X = x$ (**mean response**) or whether one is predicting the actual Y for a particular individual (** single observation**). For this assignment, we will concentrate on the confidence interval for the mean response. We do so because it is rare to use statistical models in public health as forecasting models (predicting an

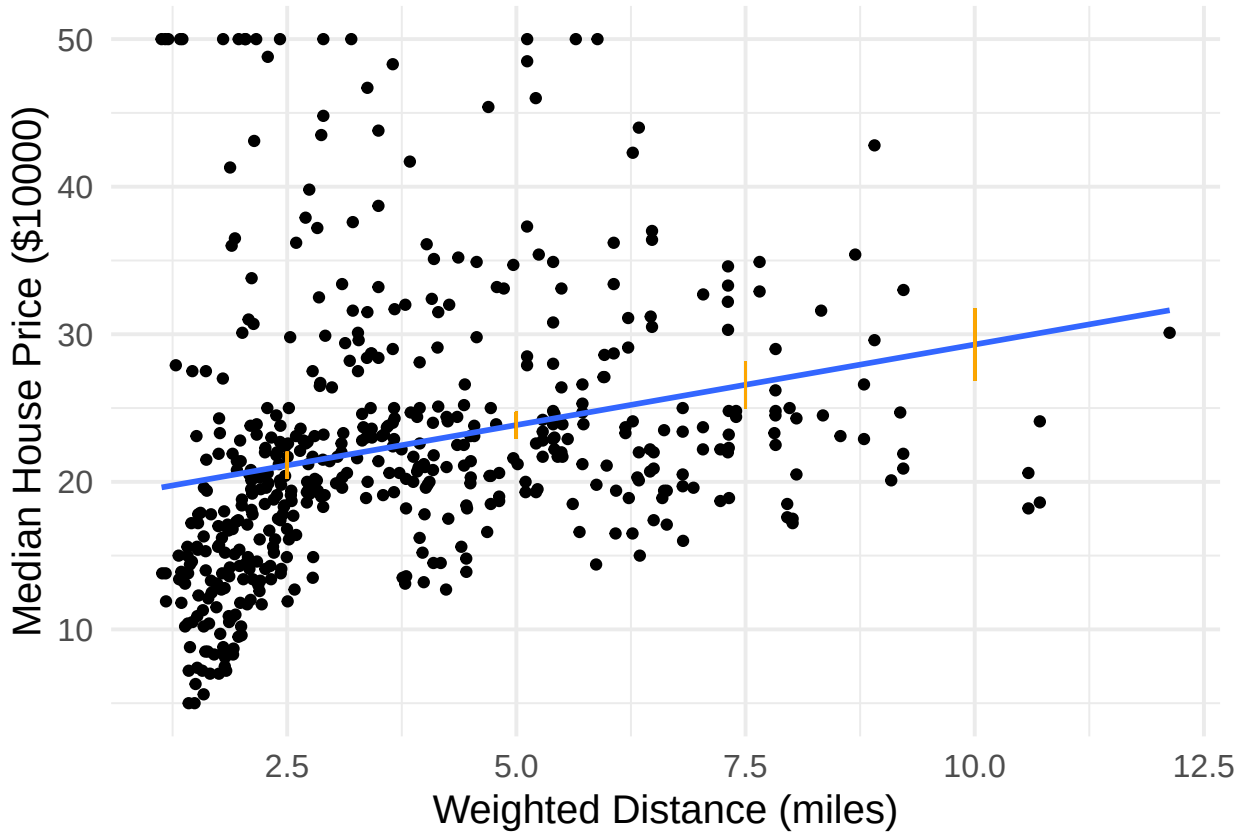
individual's health in the future) and more common to use them to estimate population-level changes (how does the mean health change in a population as we change exposure). However, as precision medicine becomes more of a reality and the models accurately predict health (i.e., have high R^2's), then statistical forecasting may become more common in our field

7. Calculate four 95% confidence intervals, one at each `dis` value: 2.5, 5.0, 7.5, and 10.0 miles. Add the four CIs to a scatter plot of the data (along with the line of best fit):

```
#Put your code here

#solution
#(note there are many ways to do this. YOu can follow the method from class and
#calculate the interval separately for each value of `dis`)
atx <- c(2.5, 5.0, 7.5, 10.0)
CIs <- predict(lm_dis, newdata = data.frame(dis=atx), interval = "confidence")
CIs
```

```
##        fit      lwr      upr
## 1 21.11912 20.20485 22.03339
## 2 23.84815 22.95092 24.74539
## 3 26.57719 25.00035 28.15402
## 4 29.30622 26.88135 31.73108
```

```
ggplot(dat, aes(x = dis, y = medv)) +
  theme_minimal(base_size = 15) +geom_point()+
  labs(x = "Weighted Distance (miles)", y = "Median House Price ($10000)")+
  geom_smooth(method = "lm", se = F)+ geom_segment(aes(x = atx[1], xend = atx[1], y = CIs[1,2], yend = C
geom_segment(aes(x = atx[2], xend = atx[2], y = CIs[2,2], yend = CIs[2,3]), col = "orange", alpha = 0.4)
  geom_segment(aes(x = atx[3], xend = atx[3], y = CIs[3,2], yend = CIs[3,3]), col = "orange", alpha = 0
geom_segment(aes(x = atx[4], xend = atx[4], y = CIs[4,2], yend = CIs[4,3]), col = "orange", alpha = 0.4)
```



8. Interpret the pointwise 95% of the median house price when distance = 10.

<Your answer here.>

Solution: The 95% CI for the average household value when distance is equal to 10 miles goes from 21.77 to 45.46 (or $21,770 to $45,460 since the response variable is in 1000's). We used a method that 95 times out of 100 the interval we produced will contain the true value of the average response variable (median house price)

9. Why do you think the CI's get wider as the `dis` gets larger?

<Your answer here.>

Solution: Because there are fewer data points as distance increases.

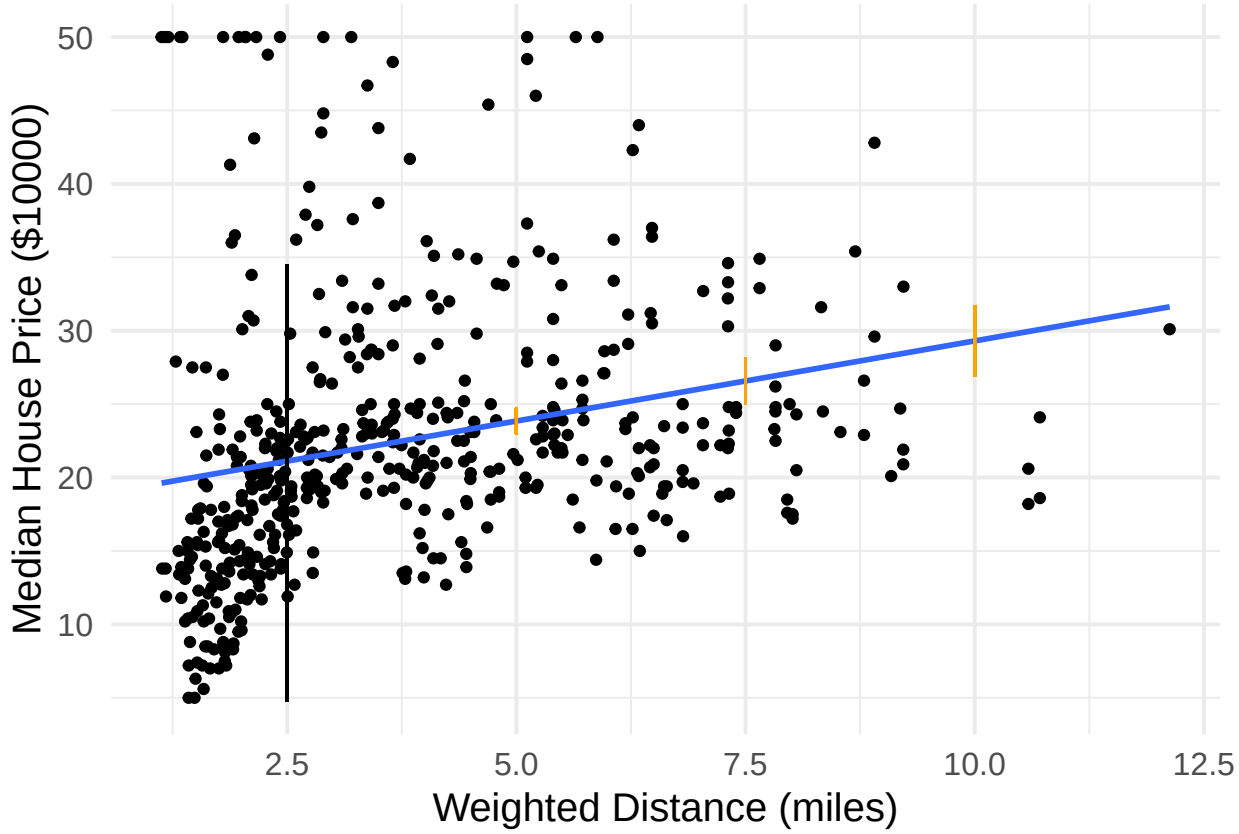
10. Add a prediction interval at x = 2.5. Why is is wider or more narrow?

```
#Put your code here

#solution
PIs <- predict(lm_dis, newdata = data.frame(dis=atx), interval = "predict")
PIs
```

```
##        fit       lwr      upr
## 1 21.11912  3.581985 38.65626
## 2 23.84815  6.311897 41.38441
## 3 26.57719  8.993056 44.16132
## 4 29.30622 11.625856 46.98658
```

```
ggplot(dat, aes(x = dis, y = medv)) +
  theme_minimal(base_size = 15) +geom_point()+
  labs(x = "Weighted Distance (miles)", y = "Median House Price ($10000)")+
```

```
  geom_smooth(method = "lm", se = F) +
  geom_segment(aes(x = atx[1], xend = atx[1], y = CIs[1,2], yend = CIs[1,3]), col = "orange", alpha = 0
  geom_segment(aes(x = atx[2], xend = atx[2], y = CIs[2,2], yend = CIs[2,3]), col = "orange", alpha = 0
  geom_segment(aes(x = atx[3], xend = atx[3], y = CIs[3,2], yend = CIs[3,3]), col = "orange", alpha = 0
  geom_segment(aes(x = atx[4], xend = atx[4], y = CIs[4,2], yend = CIs[4,3]), col = "orange", alpha = 0
  geom_segment(aes(x = 2.5, xend = 2.5, y = 4.773799, yend = 34.47926))
```



<Your answer here.>

Solution: Wider because 95 times out of 100, the CI will contain any individual's value when `dis=2.5`, whereas the CI needs to just contain the average measure of price, and the average is less variable then the mean.