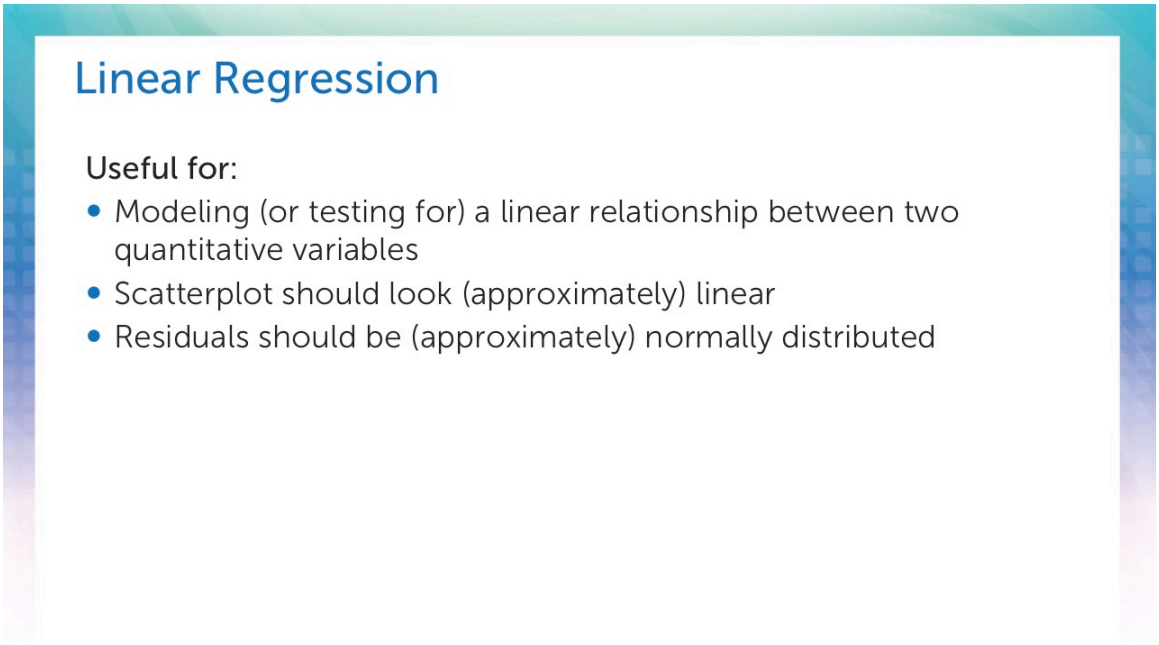

DS 710
Programming for Data Science
Linear Regression in R

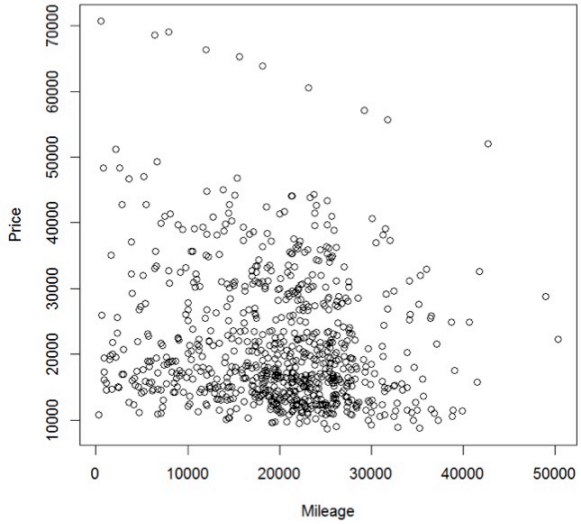


Linear regression is useful when you're modeling or testing for a linear relationship between two quantitative variables. To use linear regression, the scatterplot should look approximately linear and the residuals should be approximately normally distributed. We'll talk more about this second condition later.

Example

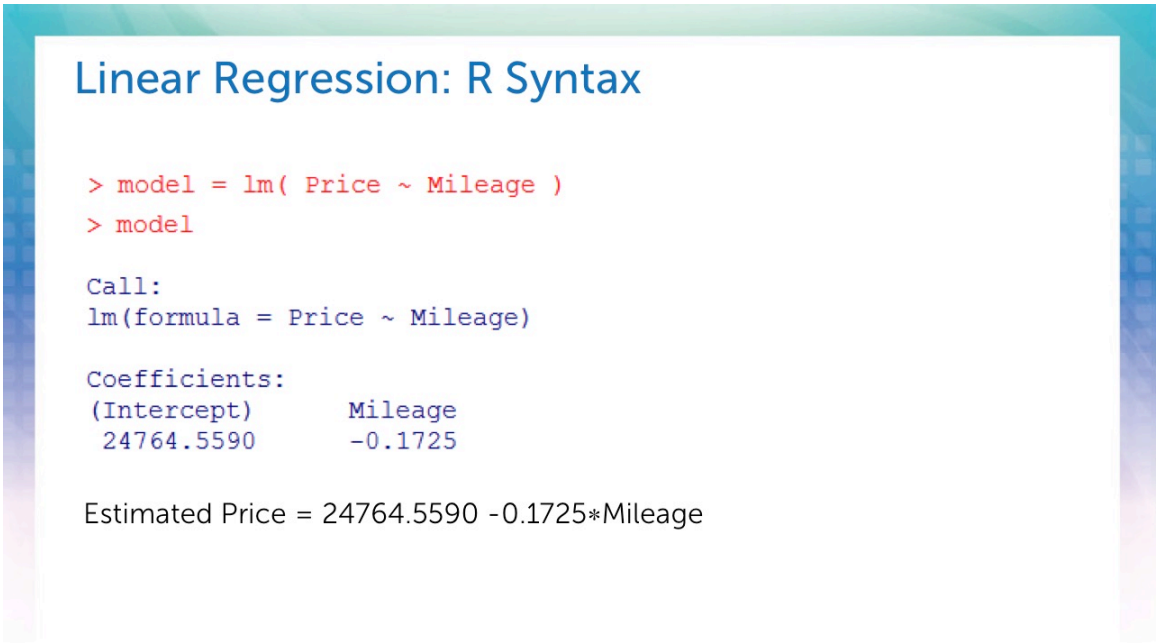
Cars 2005.csv

```
> plot(Mileage, Price)
```



For example, suppose we want to find a model for the relationship between price and mileage for used cars. We'll use the car's 2005 data set for this example. So we're focusing on one-year-old used cars in 2005.

It's always a good idea to look at your data before you get started. So we can do a scatterplot between mileage and price using the plot function. And we see it looks something like this. Here, the relationship isn't super linear. We can't see a straight line in the data. But we also don't see any curved pattern that would make us think that a linear model wouldn't be appropriate.

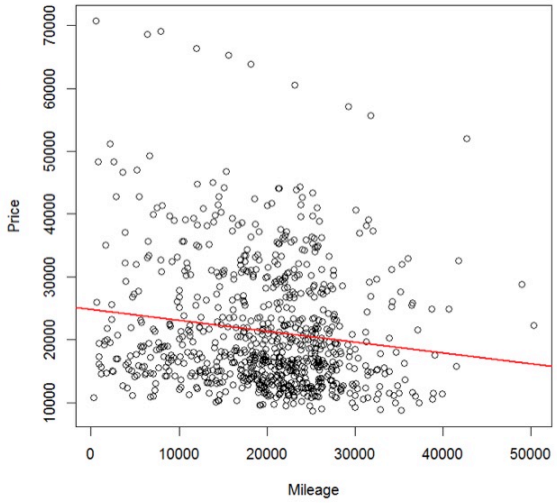


To do the linear regression in R, we can use the LM function, which stands for linear model. The syntax here is a variable name-- I called it model-- equals LM parentheses the response variable or y variable-- price-- tilde, the predictor variable or x variable-- mileage. When you type this into R, nothing will show up on the screen right away. To see the results, you need to type model again and press Enter.

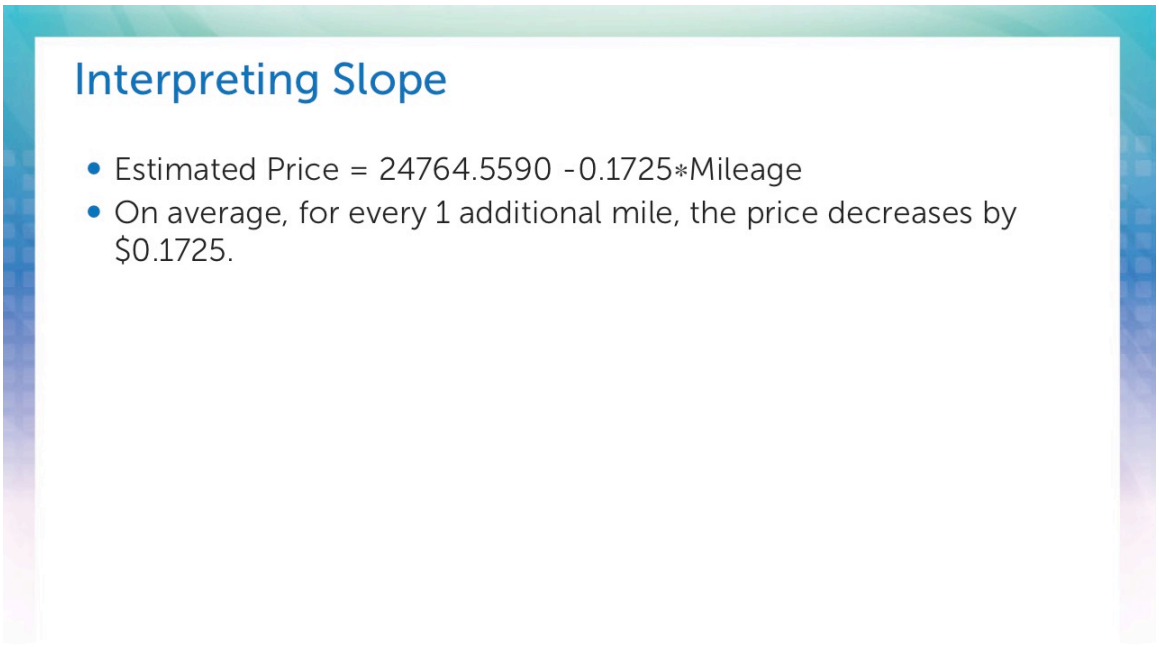
So here we can see that the variable model contains a model with two coefficients, a y-intercept and a slope for the mileage variable. This means that the equation for our model looks like this. The estimated price for a car is $24,764 minus 0.17 times whatever the mileage of that car is.

Plotting the Best-fitting Line

```
> plot(Mileage, Price)
> abline( model, col = "red", lwd = 2 )
```



This is the equation of the best fitting line for our data set. And we can see this line plotted versus the data using the AB line function. So AB line of our linear model. And here I've chosen to make the color red and the line width 2. This will add the line to the scatterplot, so you need to make sure your scatterplot is already showing in the graphing window in R.



An important thing to do when you have a linear model is to be able to interpret the meaning of the slope. In this case, our interpretation is that on average for every one additional mile, a car's price decreases by $0.17.

SELF-ASSESSMENT: MULTIPLE CHOICES

The R output shown here gives the results of a linear regression of people's blood alcohol content (BAC) on the number of beers they had drunk. Which of the following interpretations is most appropriate?

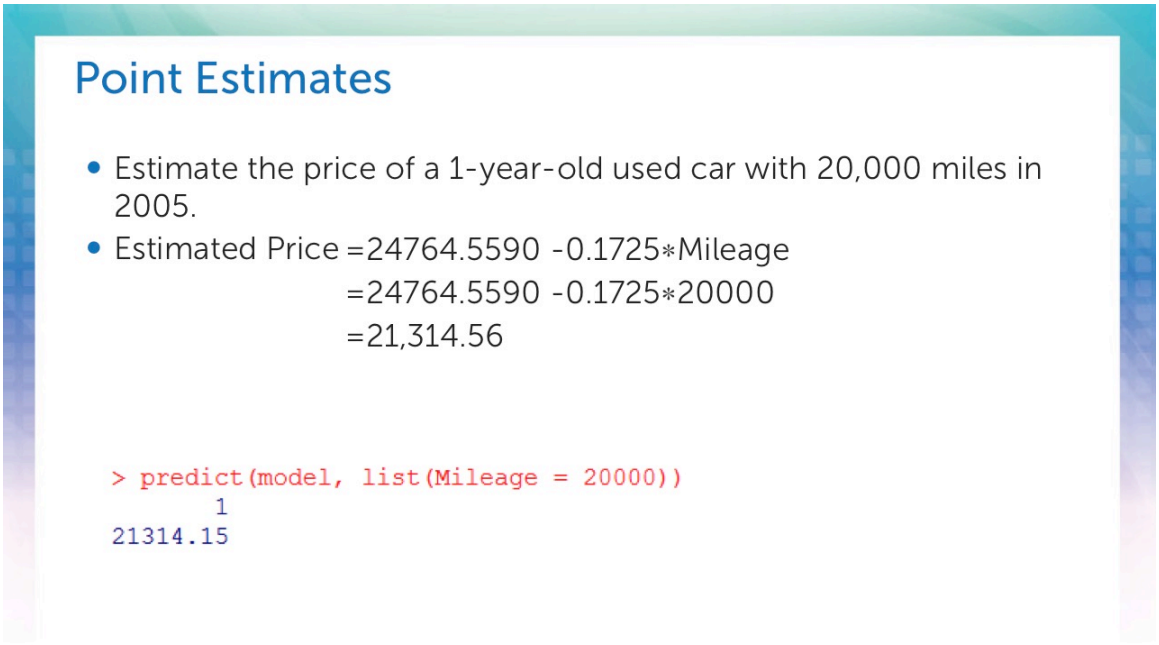
```
> model = lm( BAC ~ Beers )
> model

Call:
lm(formula = BAC ~ Beers)

Coefficients:
(Intercept)        Beers
   -0.01270      0.01796
```

- For every 1 additional beer a person drinks, we expect his or her BAC to increase by 0.01270.
- An increase in BAC of 1 is associated with drinking 0.01796 more beers.
- For every 1 additional beer a person drinks, his or her BAC increases by 0.01796.
- For every 1 additional beer a person drinks, we expect his or her BAC to increase by 0.01796.

SUBMIT

Question 1

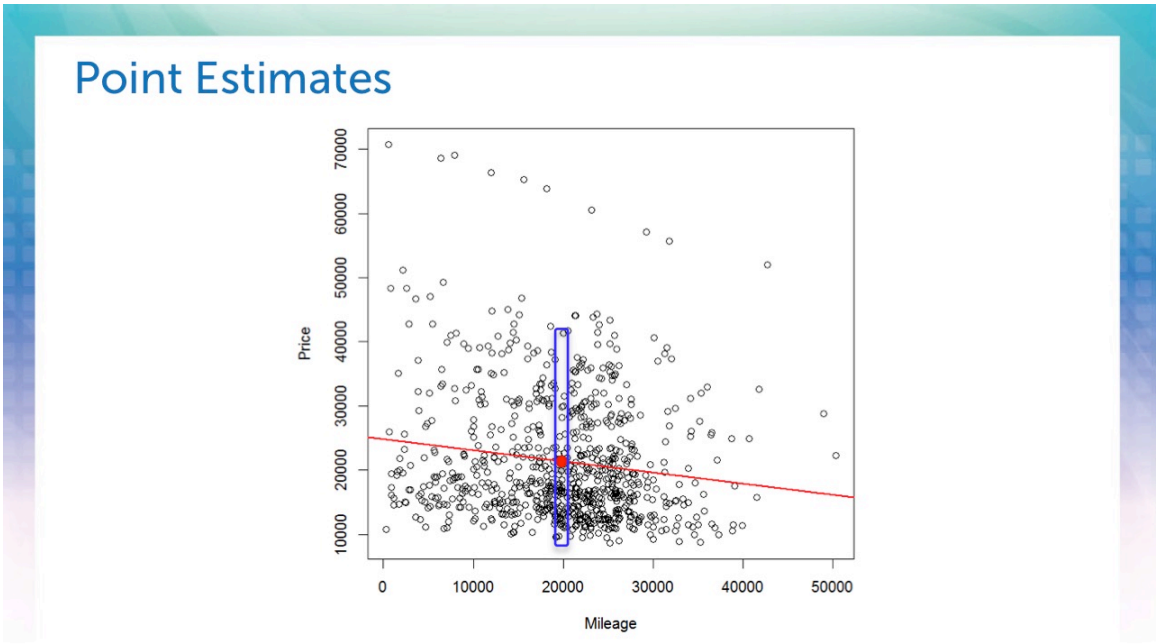
See correct answer at the end of the transcripts.



Suppose we want to estimate the price of a one-year-old used car with 20,000 miles on it in 2005. We can do this using our equation for the model by plugging in 20,000 for the mileage. And we get an estimated price of $21,314. This is a point estimate because we're getting a single point value for the estimated price, not an interval or range of plausible values.

We can also do this calculation in R using the predict function. The first argument is the variable that stores our linear model. So in this case, model. And the second argument should be a list containing the values of all of the predictor variables. In this case, we just had one predictor variable, mileage.

You can see that the value of this prediction in R is slightly different than the one we got by hand. It's off by a few cents. And it that's due to rounding errors, that R didn't show us the full decimal place accuracy of the estimate for the y-intercept and slope.



What we're doing in finding the point estimate is looking at that line of best fit shown in red and finding the height of the point on that line that had an x value of 20,000. And the y value, or height of that point, is our estimated price for that car. You can see that there are quite a few cars with mileages near 20,000 miles and they had a wide range of prices.

So our estimate here is just that. It's just an estimate that doesn't take into account random noise or all of the other factors that might affect the price of a car.

SELF-ASSESSMENT: MULTIPLE CHOICES

The R output shown here gives the results of a linear regression of people's blood alcohol content (BAC) on the number of beers they had drunk. Find a point estimate of the BAC of a person who drank 2 beers.

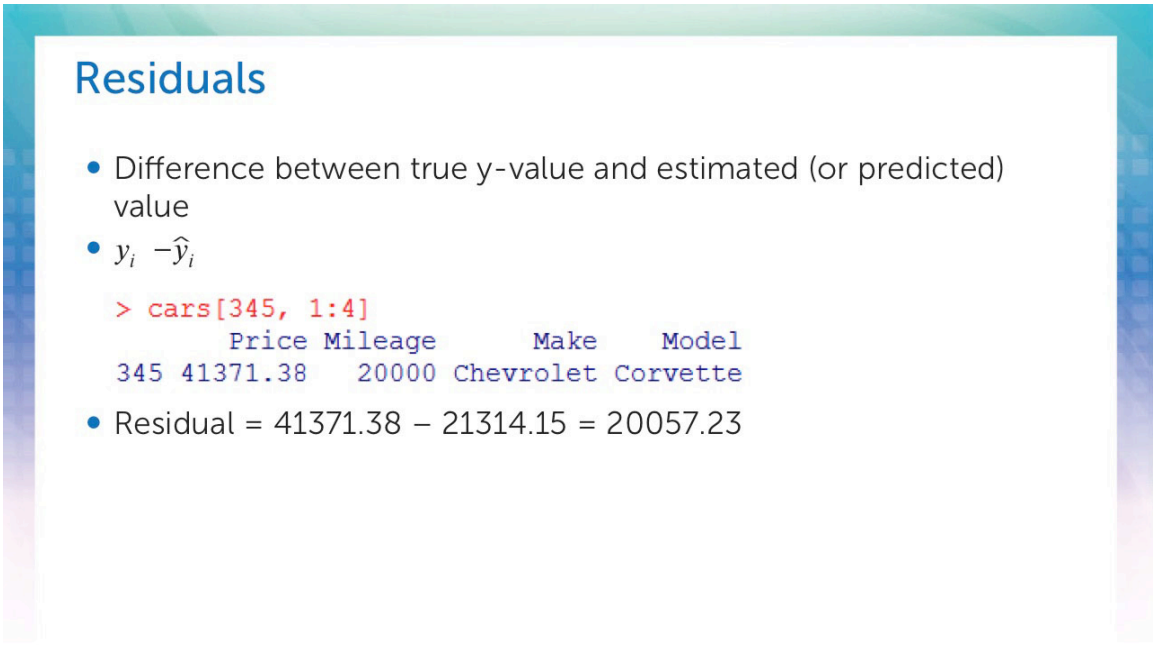
```
> model = lm( BAC ~ Beers )
> model

Call:
lm(formula = BAC ~ Beers)

Coefficients:
(Intercept)        Beers
   -0.01270      0.01796
```

- -.00744
- .02322
- .01052
- .04862

SUBMIT

Question 2

See correct answer at the end of the transcripts.



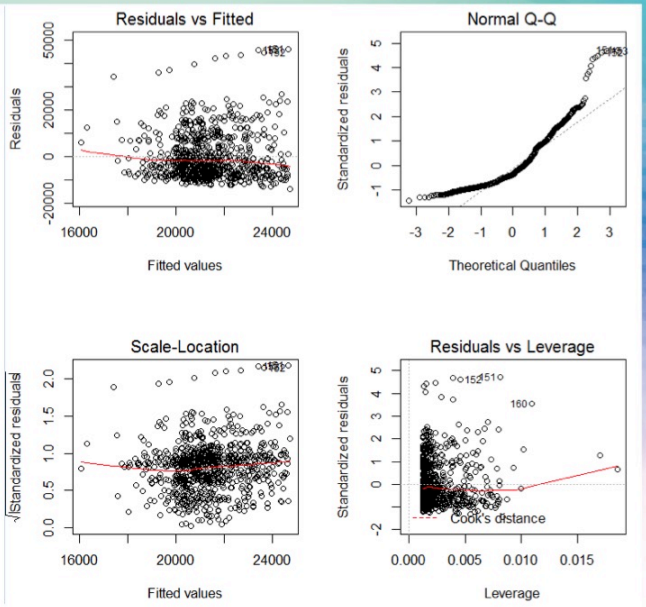
The residual of a point is the difference between it's true y value observed in the data and its estimated or predicted y value. This is often written as y sub i for the true y value of the i-th point in the data minus y sub i hat, where the hat represents an estimate or prediction.

For example, in our cars data set is 345th row actually contains a car with a mileage of exactly 20,000. And it had a price of $41,371, which is quite a bit higher than what we predicted for a car with a mileage of 20,000. So the residual of this point is the difference between it's true price in the data and the predicted price based on its mileage.

Evaluating Appropriateness of Linear Regression

Residual diagnostic plots

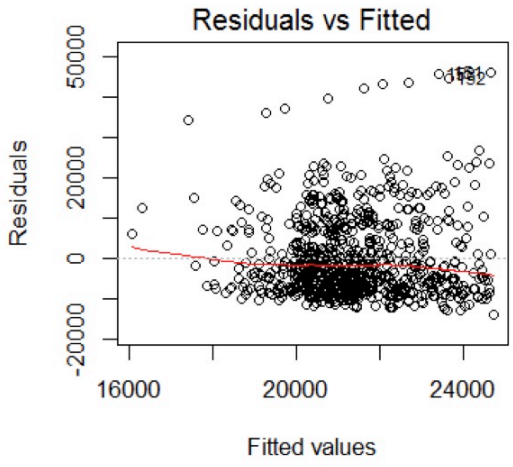
```
> par( mfrow = c(2,2) )
> plot( model )
```



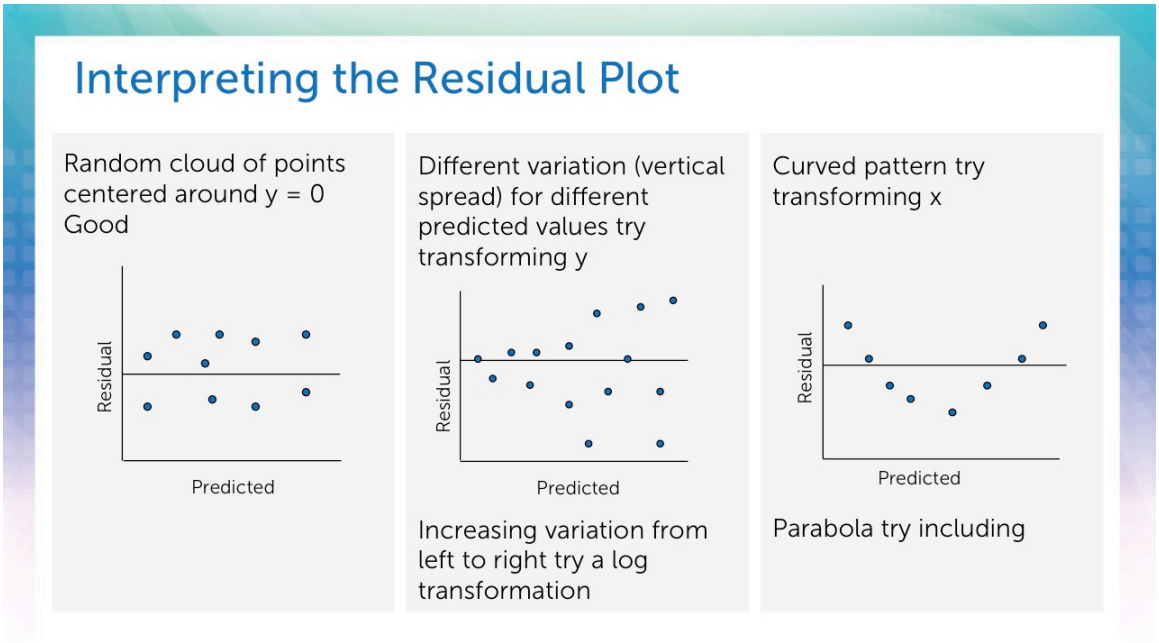
We can use residuals to help us evaluate the appropriateness of doing linear regression. And we do this using the residual diagnostic plots. To look at these, we start by using the PAR function with MF row equal to a vector with two rows and two columns to prepare our graphing window to plot four graphs at once.

Then we can simply use plot and the variable name where we stored the linear model to see the diagnostic plots. And they'll look something like this. For our purposes, we mainly want to focus on the top two plots.

Residual vs. Fitted Plot



The plot in the upper left-hand corner is residuals versus fitted values plot. So the y-axis here shows the residuals. So each point's y value minus it's predicted y value. And the x-axis shows the fitted values. That is, the predicted y values.

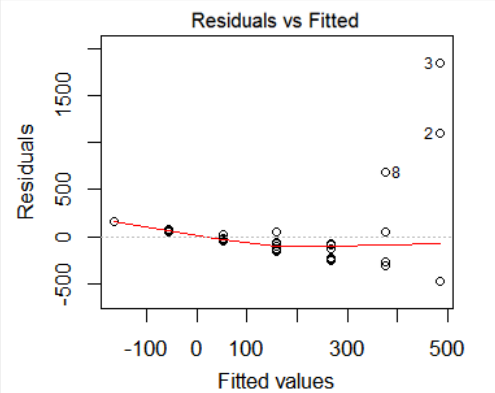


To interpret this residual plot, you want to see a random cloud of points centered around the line y equals 0. That's a good sign that the residuals are random. There's no trend in the data that's not being accounted for by the linear regression.

An example of something you don't want to see in the residual plot is changes in variation, meaning changes in the amount of vertical spread in the data for different predicted values. If you see this kind of pattern, it's a good idea to try a transformation on the y variable. In particular, if you see increasing variation from left to right on the residual plot, as shown in this picture, it's a good idea to try a log transformation.

And finally, if you see a curved pattern to the residuals it's a good idea to try transforming the x variable. And in particular, if you see a parabola it's a good idea to try including an x squared term in your regression model.

The residual plot shown here is from a linear regression of the amount of time (y) until an insulating substance broke down, as a function of the electric voltage (x) it was exposed to. What should be the next step in this analysis?



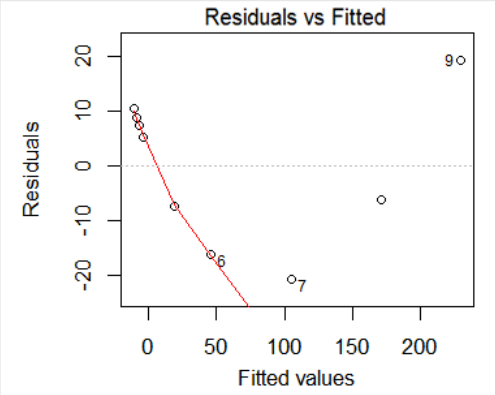
- Write the analysis report; linear regression is appropriate here.
- Try taking the natural log of the breakdown times.
- Try squaring the voltage.

SUBMIT

Question 3

See correct answer at the end of the transcripts.

The residual plot shown here is from a linear regression of the length of a year (y) on different planets, as a function of their distance (x) from the sun. What should be the next step in this analysis?



- Write the analysis report; linear regression is appropriate here.
- Try taking the natural log of YearLength.
- Try including Distance² in the regression model.

SUBMIT

Question 4

See correct answer at the end of the transcripts.

Including $x\uparrow 2$ in the Model

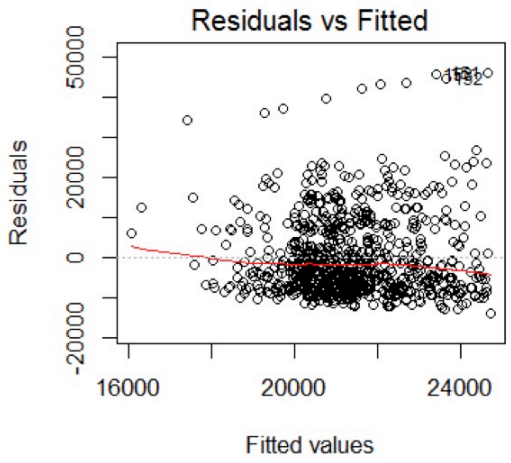
```
> Distance2 = Distance^2
> model2 = lm( YearLength ~ Distance + Distance2 )

> model2

Call:
lm(formula = YearLength ~ Distance + Distance2)

Coefficients:
(Intercept)     Distance    Distance2
 -2.230e+00    2.854e-02    1.087e-05
```

YearLength = -2.23 + 0.02854 * Dist + 0.00001087 * Dist↑2

In the previous quiz question, we decided it would be a good idea to try including an x squared term in our model. We can do this by creating a new variable called distance 2 that's the square of the variable distance. Then our linear model, which I'll call model 2, will involve the year length as a function of distance and our new variable distance 2. And we use a plus sign represent and.

If we then look at model 2, we can see that we get three coefficients-- a y-intercept and a coefficient that's multiplied by distance and by distance squared. This would represent the equation year length equals negative 2.23 plus 0.02854 times the distance plus 1.087 times 10 to the negative 5 times the distance squared.

Residual Plot for Price vs. Mileage

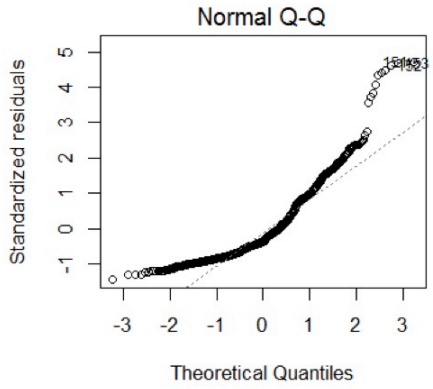


Let's go back to our price versus mileage model. The residual plot here is somewhat ambiguous. Some of the residuals are very far from 0, as you can see on the y-axis of this plot. But there doesn't appear to be a strong trend of either a curve or a change in variance or vertical spread in this residual plot.

There might be a slight tendency of increasing variance as we go from left to right here, but it's not nearly as pronounced as in the example you saw in the quiz question. So overall, based on this residual plot, I might want to try doing a transformation. But if that didn't seem to improve the fit of the model, I might be OK with simply using the linear regression, as we've done here.

Normal Quantile-Quantile Plot

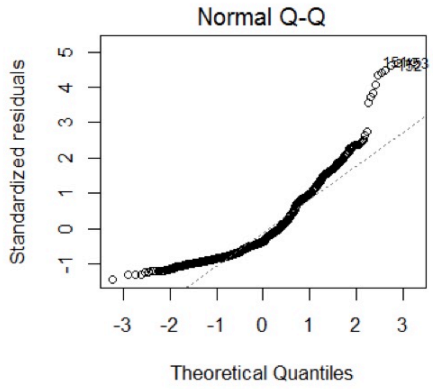
How well do the residuals fit a normal distribution?



We can get more information about whether linear regression is appropriate using the normal quantile quantile plot, or QQ plot for short. This is the second of our diagnostic plots, and it's the one that shows up in the upper right-hand corner. This plot helps us answer the question, how well do the residuals fit a normal distribution? Remember that to use linear regression, the residuals should be approximately normally distributed.

Interpreting the Normal Quantile-Quantile Plot

How well do the residuals fit a normal distribution?

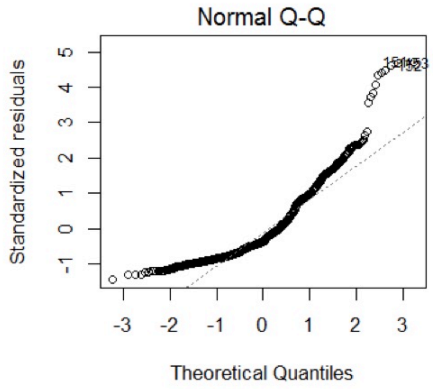


So what does the normal quantile quantile plot show us? Well, the y-axis shows the standardized residuals. That is, we take the residuals from the data, subtract the mean, and divide by the standard deviation, creating a z-score for each residual. The x-axis shows the theoretical quantiles, or the z-scores that we would expect if we had a data set of the same size as this one but one that was perfectly normally distributed.

So if the residuals fit a normal distribution, then we expect each of the standardized residuals to line up with the same value on the theoretical quantiles. 0 would line up with 0, 1 would line up with 1, and so on. So if the residuals are approximately normally distributed, then the points on the normal QQ plot should fall close to the line y equals x, which is shown here in the dashed gray line.

Price vs. Mileage: Normal Quantile-Quantile Plot

How well do the residuals fit a normal distribution?

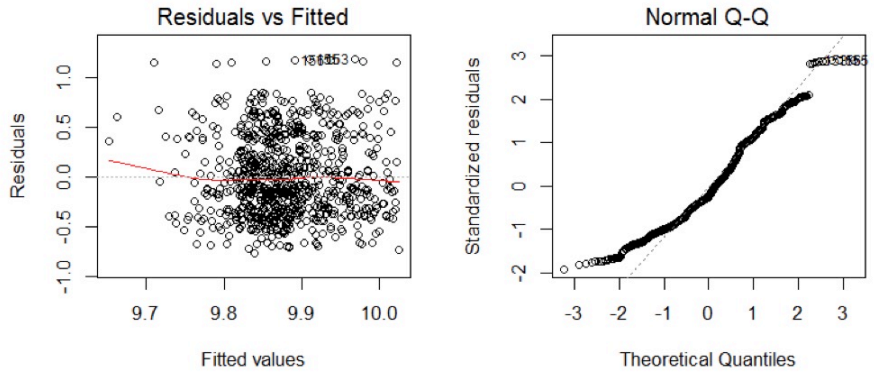


In our price versus mileage analysis, the residuals don't fall that close to the line y equals x, especially when you look in the upper right-hand corner of the normal QQ plot. That makes me think that linear regression might not be the most appropriate for this data set and we might want to try a transformation to get a better fitting model.

Remember that on the residuals versus fitted plot, we saw a slight hint of increasing variation or vertical spread as we move from left to right. So I'll try a log transformation.

A Better Model for Price and Mileage

```
> logPrice = log(Price)
> model2 = lm(logPrice ~ Mileage)
> plot(model2)
```



We can log transform price simply by creating a new variable called log price using the function log. Remember that log in R stands for the natural log. Then we can create a new model, model 2, using LM of log price tilde mileage.

When we look at the residual plots for this model, we see that the residuals overall are much smaller. They range from about negative 1 to positive 1 instead of negative 20,000 to positive 20,000. And we don't see as much of a trend of increasing variation from left to right.

The normal QQ plot still isn't great, especially in the upper right- and lower left-hand corners, but it looks somewhat better. So let's proceed using this log transformed model.

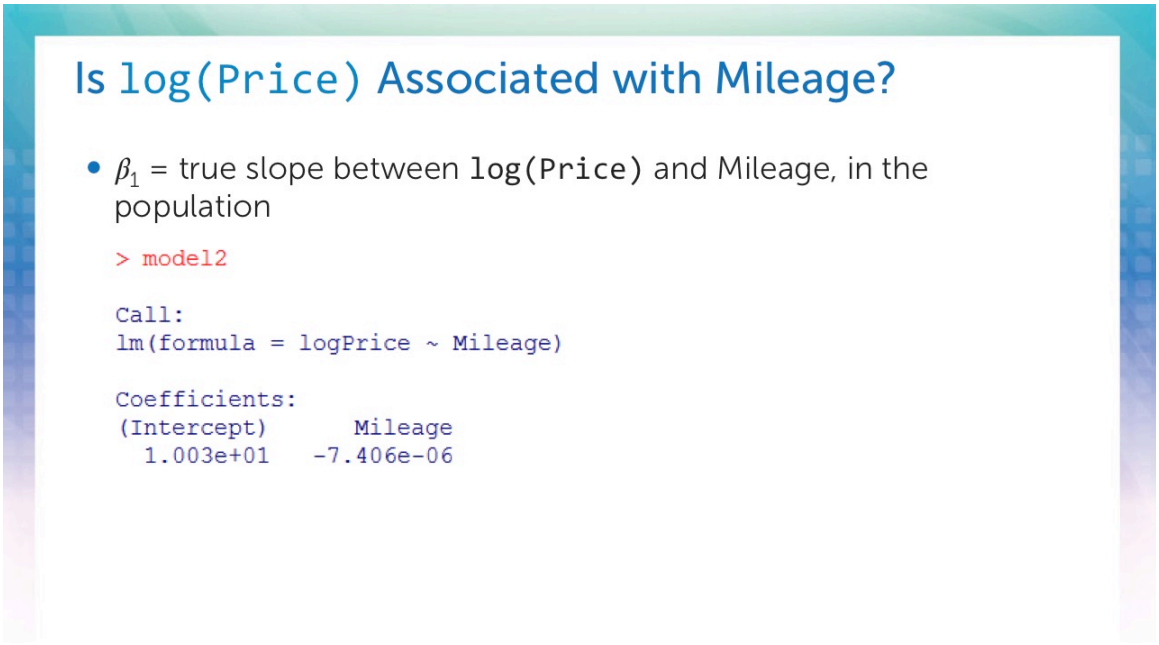
The Model Equation

```
> model2

Call:
lm(formula = logPrice ~ Mileage)

Coefficients:
(Intercept)      Mileage
  1.003e+01   -7.406e-06
```

Estimated$\log(Price) = 10.03 - 7.406 * 10^{-6} * Mileage$

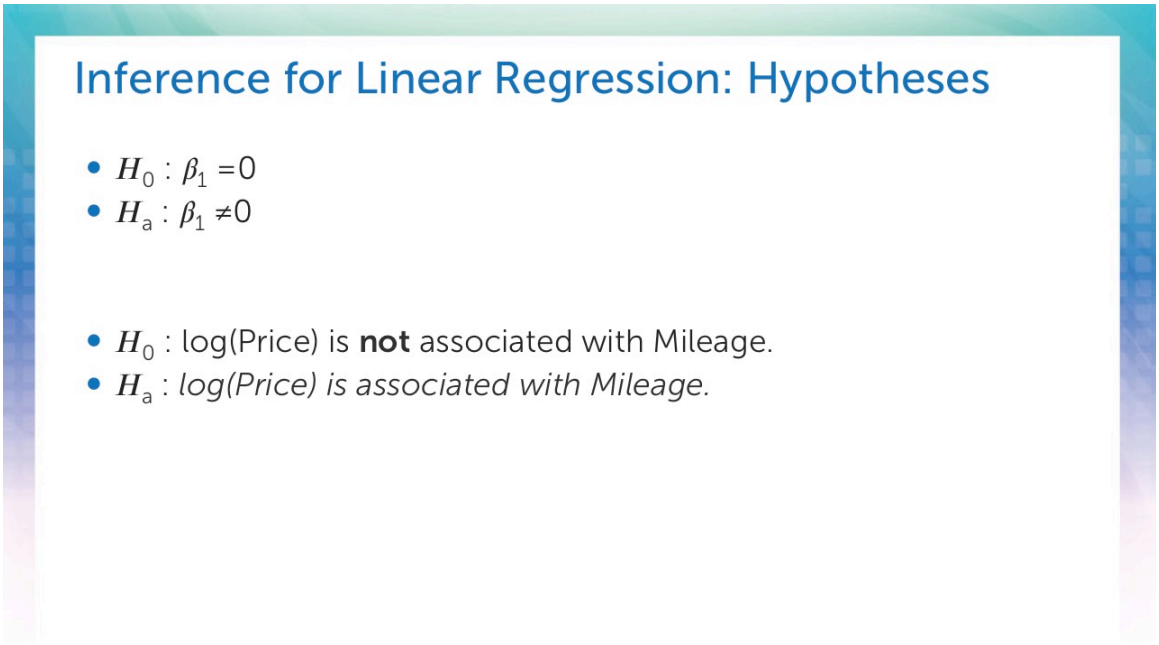
Estimated$\log Price = e^{10.03 - 7.406 * 10^{-6}} * Mileage$

When we look at model 2, we again get a y-intercept and a slope. But the interpretation is slightly different. Here our equation is for the estimated log price rather than the estimated price. We could turn this into an equation for the estimated price by taking e to both sides of this equation. e raised to the natural log just cancels out, and we're left with price on the left-hand side of this equation, and on the right-hand side we get e raised to the y-intercept plus the slope times the mileage.



Now let's try to generalize these results and find out whether log price is associated with mileage in general in a larger population. Here we're interested in beta sub 1, the true slope between log price and mileage if we could measure every single car in the population of one-year-old used cars in 2005. And we want to know whether this slope is different from 0 or not.

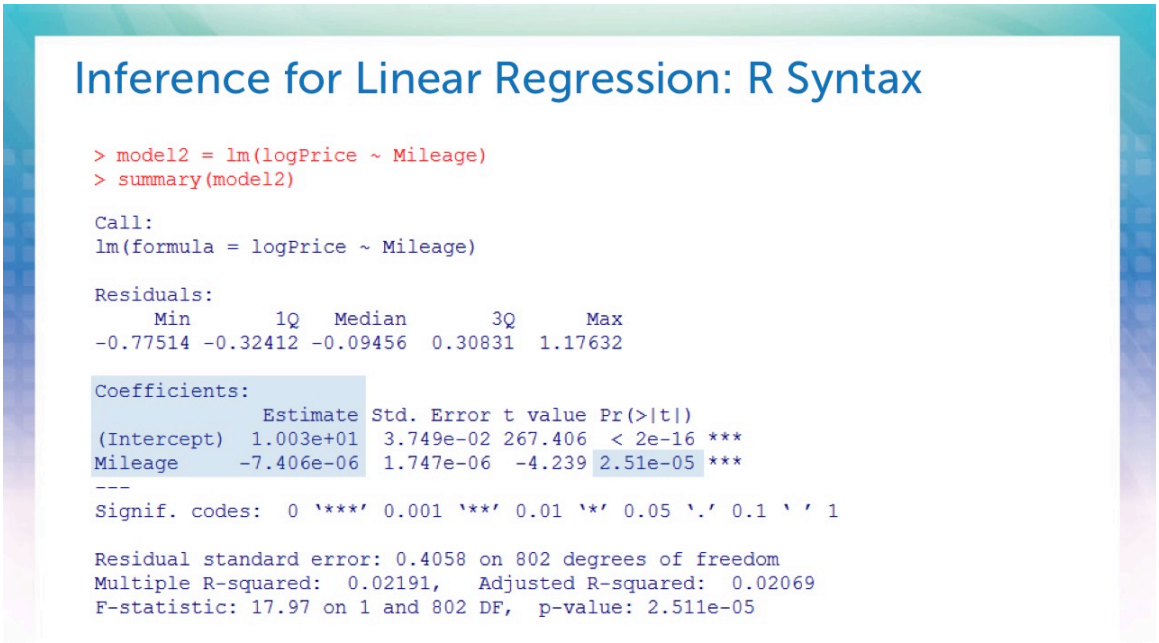
We've already seen that in our sample the slope between log price and mileage was negative 7.406 times 10 to the negative 6. That is different from 0, but it's a pretty small number pretty close to 0. So is it different enough from 0 that we can generalize from our sample to a larger population?



To do inference for linear regression, our null hypothesis is that beta 1, this true slope in the population, is equal to 0. And our alternative hypothesis is that beta 1 does not equal 0.

Now if you think about the interpretation of this slope, if the null hypothesis were true, if the slope really were 0, that would mean that no matter what mileage a car had, the estimated price, or in this case estimated log price, would be the same. So there would be no relationship between mileage and log price.

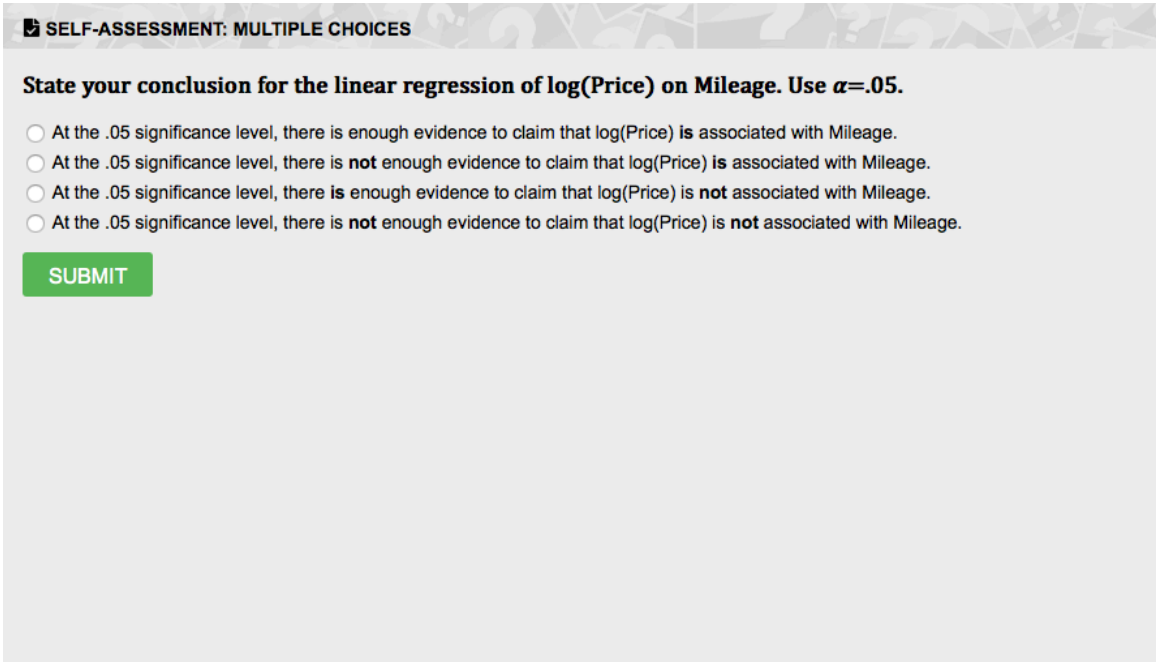
In other words, the null hypothesis is that log price is not associated with mileage. And the alternative hypothesis is that log price is associated with mileage.



To do this hypothesis test in R, we're going to use the LM function, just like we did before when we were looking at the equation for the linear model and the diagnostic plots. In fact, if you already have model 2 stored in memory from looking at the diagnostic plots, you don't need to recreate it. You just need to look at the summary of model 2.

And here's what that looks like. There's a lot of information here, but we're going to focus on two pieces. First, in the section labeled coefficients, you'll notice that the first column contains estimates of the intercept-- that's the y-intercept-- and the slope for mileage. Those are the same values that we got when we just typed model 2 and pressed Enter.

The second thing to focus on here is this entry in the fourth column of our coefficients section in the second row, that is the row associated with mileage. This is the p value for whether the slope of mileage is different from 0.



Question 5

See correct answer at the end of the transcripts.

Extracting Information From summary()

```
> a = summary(model2)

> a$

> a$a$call          a$terms         a$residuals     a$coefficients  a$aliased
a$sigma         a$df            a$r.squared     a$adj.r.squared a$fstatistic
a$cov.unscaled

> a$coeff
                 Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)  1.002593e+01 3.749331e-02 267.405899 0.000000e+00
Mileage     -7.406335e-06 1.747372e-06  -4.238557 2.511178e-05

> a$coeff[1,1]
[1] 10.02593
> a$coeff[2,4]
[1] 2.511178e-05
```

Summary gives you a lot of information about a linear model. But sometimes you don't want to look at all of that information. You just want to focus on a few pieces. You can do that by assigning some other variable to be equal to the summary of model 2. Here I've used a.

Then, type a$ and press Tab twice. This will cause R to give you a variety of different options for different pieces of the summary that you can extract and look at individually. Here we're most interested in the coefficients section of the summary, so we can use a$ coefficients.

You cam actually use any abbreviation of coefficients as long as you type enough letters to make it distinct from all of the other options. So I've used a$ COEFF. Now we can focus on just the coefficients section that we're interested in. And what's more, a$ COEFF is now a matrix, so we can extract particular numbers from this matrix using the square bracket notation. For example, a$ COEFF square bracket 1, 1 gives the entry in the first row and first column. That's our estimated intercept for our best fitting line.

And a$ COEFF square bracket 2, 4 gives the entry in the second row and the fourth column. That was the p value for the slope. This is an especially useful technique when you want to use a for loop to do many linear regressions in a row and extract just part of the output that you want to focus on.

Question 1 answer:

SELF-ASSESSMENT FEEDBACK

CORRECT

The R output shown here gives the results of a linear regression of people's blood alcohol content (BAC) on the number of beers they had drunk. Which of the following interpretations is most appropriate?

```
> model = lm( BAC ~ Beers )
> model

Call:
lm(formula = BAC ~ Beers)

Coefficients:
(Intercept)        Beers
   -0.01270      0.01796
```

Your answer:
For every 1 additional beer a person drinks, we expect his or her BAC to increase by 0.01796.

Correct answer:
For every 1 additional beer a person drinks, we expect his or her BAC to increase by 0.01796.

Feedback:
Right! Saying that we expect BAC to increase by 0.01796 emphasizes the fact that the line of best fit is an estimated model, not a perfect fit.

Question 2 answer:

SELF-ASSESSMENT FEEDBACK

CORRECT

The R output shown here gives the results of a linear regression of people's blood alcohol content (BAC) on the number of beers they had drunk. Find a point estimate of the BAC of a person who drank 2 beers.

```
> model = lm( BAC ~ Beers )
> model

Call:
lm(formula = BAC ~ Beers)

Coefficients:
(Intercept)        Beers
   -0.01270      0.01796
```

Your answer:
.02322

Correct answer:
.02322

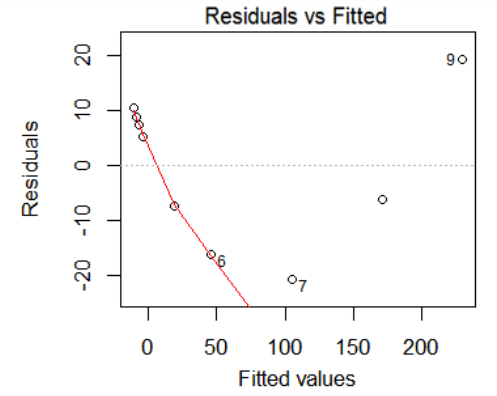
Feedback:
Great job!

Question 3 answer:

SELF-ASSESSMENT FEEDBACK

CORRECT

The residual plot shown here is from a linear regression of the length of a year (y) on different planets, as a function of their distance (x) from the sun. What should be the next step in this analysis?



Your answer:
Try including Distance² in the regression model.

Correct answer:
Try including Distance² in the regression model.

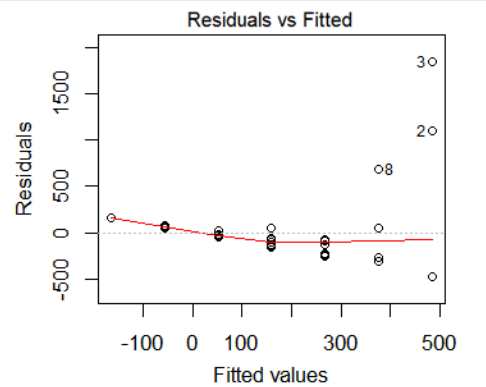
Feedback:
Yes, the residuals appear to follow a parabola.

Question 4 answer:

SELF-ASSESSMENT FEEDBACK

CORRECT

The residual plot shown here is from a linear regression of the amount of time (y) until an insulating substance broke down, as a function of the electric voltage (x) it was exposed to. What should be the next step in this analysis?



Your answer:
Try taking the natural log of the breakdown times.

Correct answer:
Try taking the natural log of the breakdown times.

Feedback:
Yes, the vertical spread between points on the residual plot increases as you go from left to right, which indicates that a log transformation might improve the fit of the model.

Question 5 answer: