DSC 152:
Applied Statistical Data Analysis and Inference

Lecture #14
Transformations

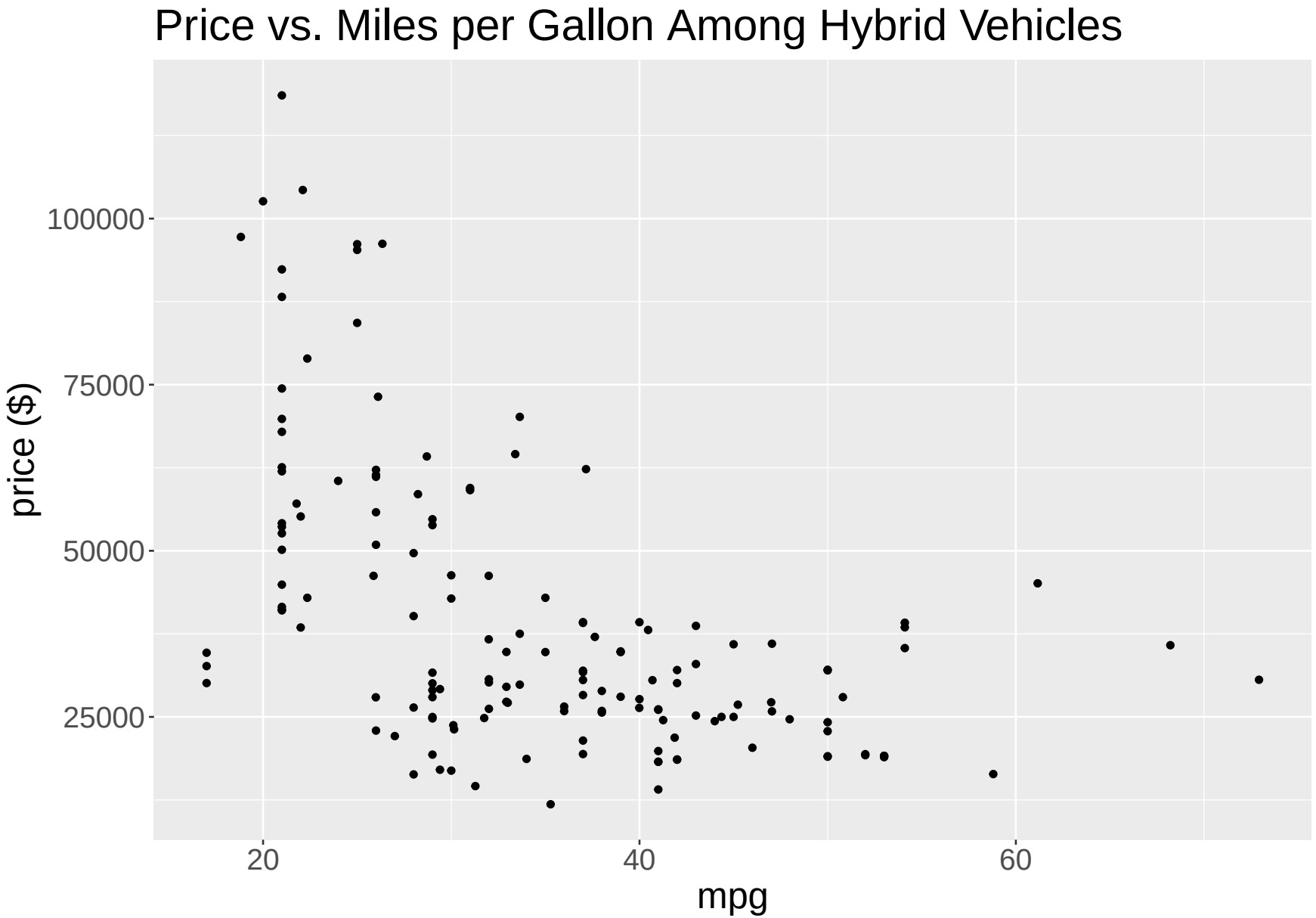
Thursday, May 14
Spring Quarter 2026
Peter Chi

Recall: Conditions for Linear Regression

- The relationship between X and Y, if there is one, is actually **L**inear
  - e.g. not quadratic, exponential, etc.
- **I**ndependence of observations
- **N**ormality of ϵ_i
  - Note that this can also be achieved, due to the central limit theorem, with a large sample size even if ϵ_i does not follow a normal distribution
- **E**qual variance across all values of X
  - Also known as homoskedasticity

What if the relationship looks like this?

A dataset from DSC 10



What if the relationship looks like this?

- The relationship is clearly not linear!
- Are we screwed?

No, there are things we can do. But...

- Everything we are going to talk about in the first part of today's lecture falls under "Sensitivity Analyses."
- In other words, hunting for the "right" transformation is a type of Model Selection, and we should not use a model selected in this manner for inference. Once we start hunting, the inference phase of our workflow is over.

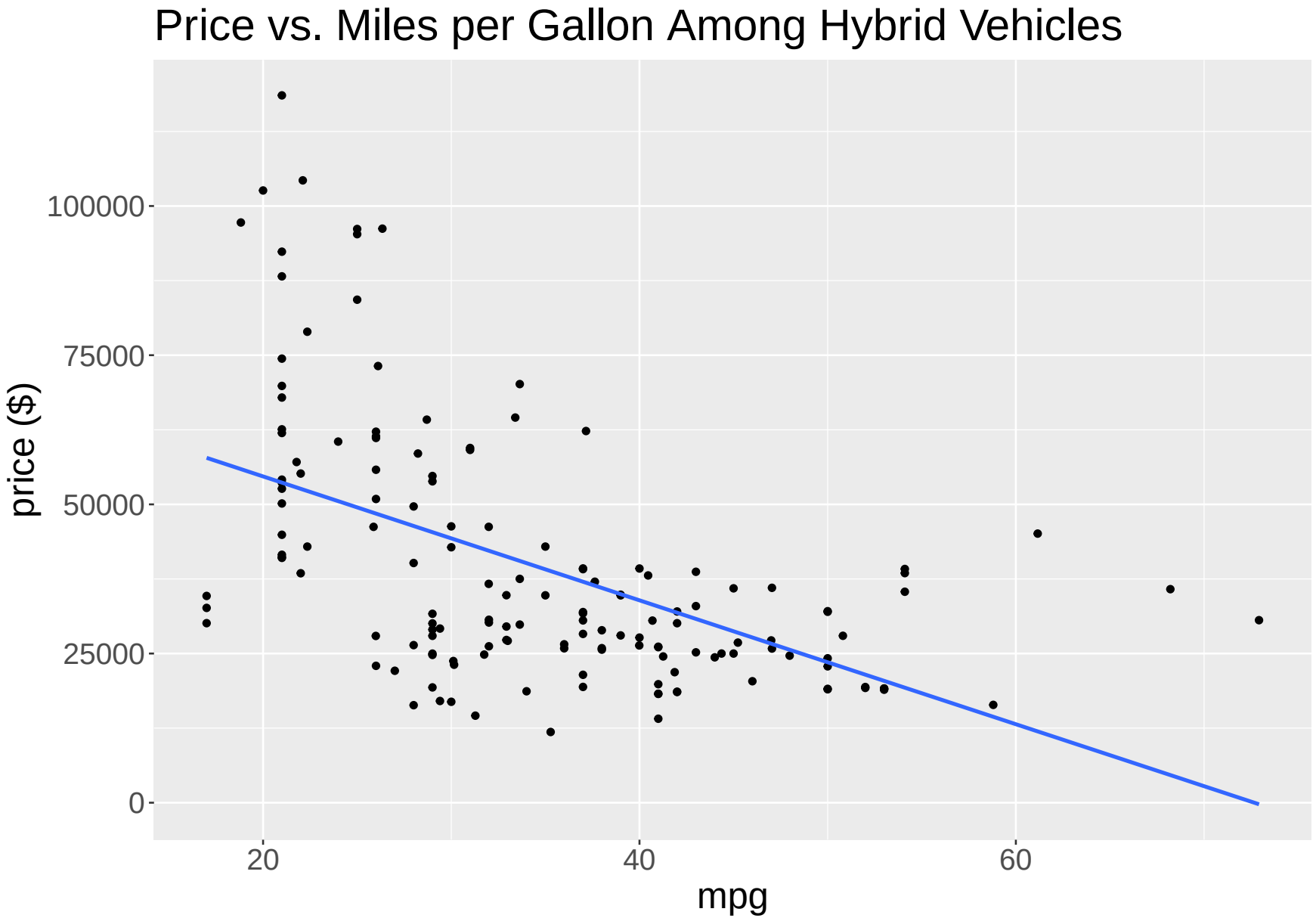
We can still run a linear model on the original data...

```r
model1 <- lm(price ~ mpg, data=hybrid)
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ mpg, data = hybrid)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -30033 -12199  -2221   8315  64899
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  75448.2     4907.5  15.374  < 2e-16 ***
## mpg          -1038.3      134.5  -7.717 1.51e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18200 on 151 degrees of freedom
## Multiple R-squared:  0.2828, Adjusted R-squared:  0.2781
## F-statistic: 59.55 on 1 and 151 DF,  p-value: 1.507e-12
```

But clearly this line is ridiculous

Anyone want a free car that gets 75 MPG??



But clearly this line is ridiculous

And what about the R^2 value?

Recall (e.g. from DSC 10 and other courses):

What is R^2?

- R^2 is a metric that indicates the degree to which the data follow a linear fit
- Its mathematical definition is:

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

where

- $\hat{y}_i$ is still just the predicted value of y according to the model
- $\overline{y}$ is the overall mean of y

Question: why is this a reasonable measure of linear fit?

But clearly this line is ridiculous

And what about the R^2 value?

More about R^2

It is NOT an absolute measure; any guidelines for what is a "good" R^2 is doomed to be inadequate, particularly as different things will happen in different situations. For example,

- In certain types of physics experiments, extremely high R^2 values are expected because e.g. gravity and friction do pretty much the same thing every time.

- In any study involving human subjects, e.g. public health, sociology, education research, etc., R^2 values will tend to be very low EVEN IF there truly is a linear trend (because people can be very different from each other, in every way imaginable).

https://jumpingrivers.github.io/datasauRus/

But clearly this line is ridiculous

What's wrong with fitting a linear model to non-linear data?

- The line will be a terrible fit at certain (if not all) values of the predictor variable (as seen on the previous slide)
- This is, to some degree, indicated by the R^2 value on the output (0.2828)
- Validity of inference is in question! (as we saw in Lab 4)

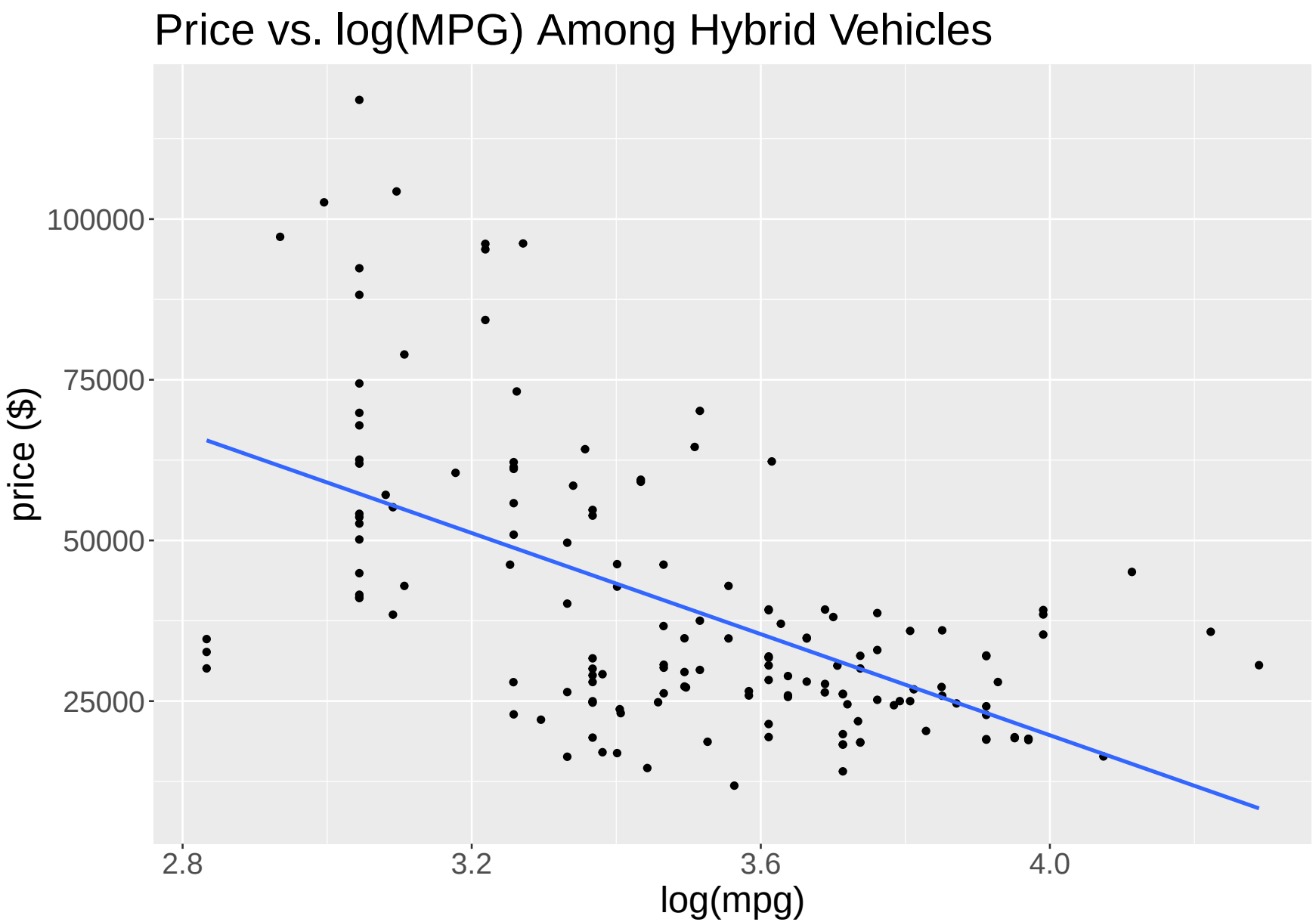
What do transformations do for us?

If we can find a function $g(x)$ such that y has a linear relationship with $g(x)$, then we can proceed with fitting:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot g(x)$$

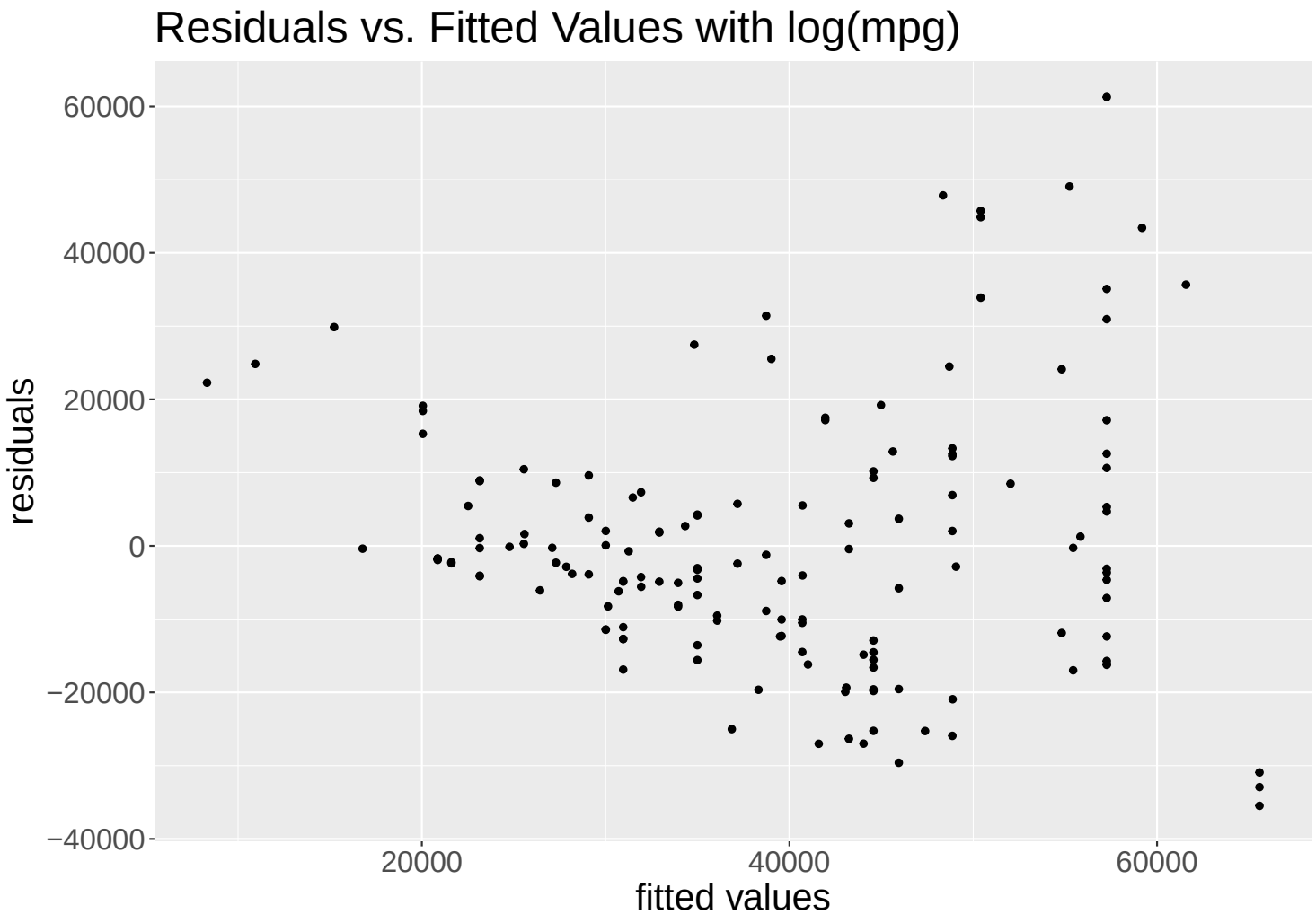
This is then still LINEAR regression (it is linear with respect to some function of x)! And all of the benefits of linear regression then apply again.

(natural) Log transformation of MPG



Log transformation of MPG

This didn't help much here. Here is the residual plot where it is pretty obvious that this is still not good:



Log transformation of MPG

And the new R^2 value:

```
model2 <- lm(price ~ log(mpg), data=hybrid)
summary(model2)$r.squared
```

```
## [1] 0.3326834
```

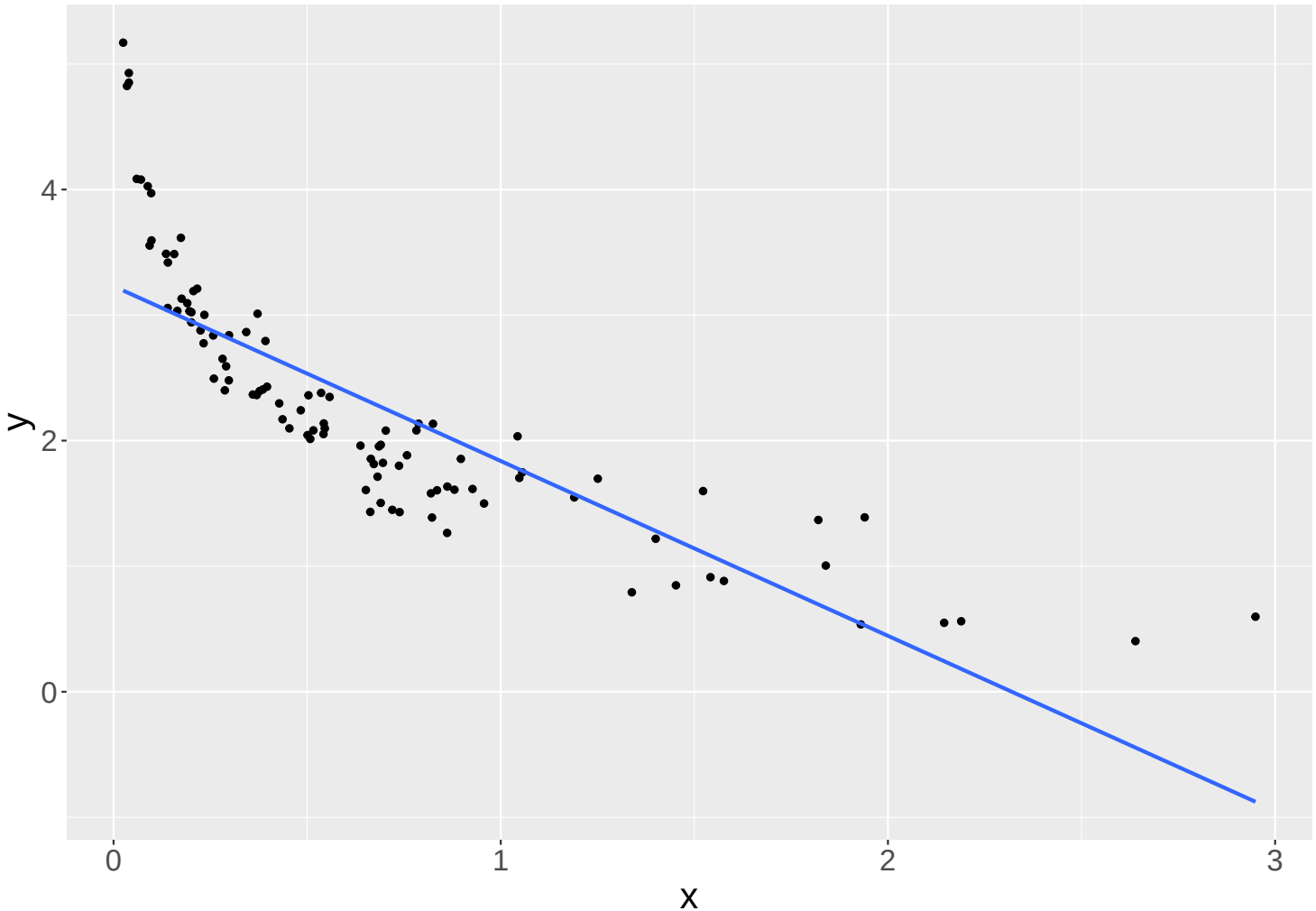
We do see some improvement from that of the untransformed model (again, 0.2828), but it's not by all that much.

Question: how much would be enough?

Again, any guidelines are not going to be universally correct. A reasonable goal might be to find the transformation that gives the greatest increase in R^2, *in combination* with an improved residual plot.

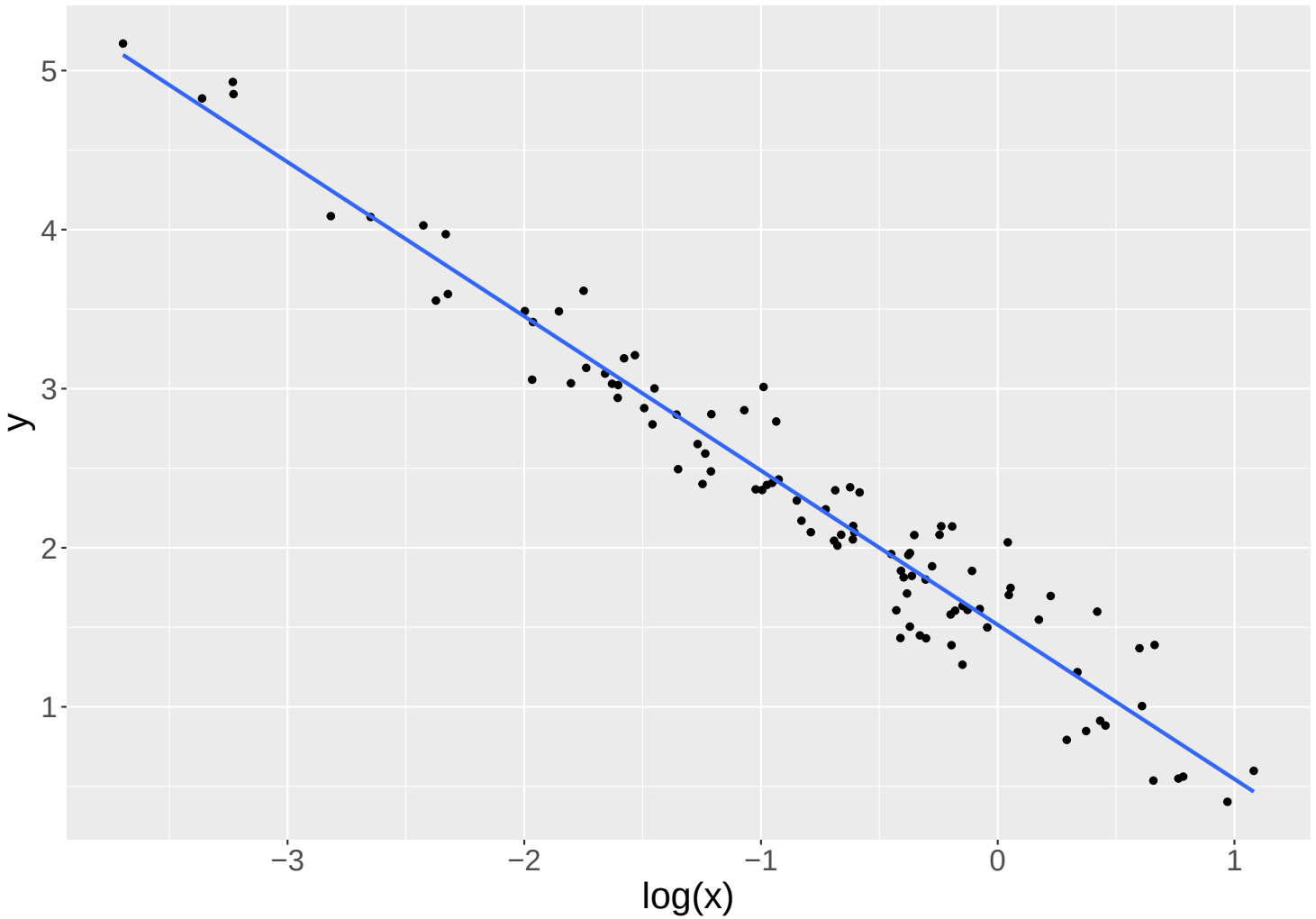
Log transformation of x

Now here is a (simulated) example where log-transforming the x-variable works really well:



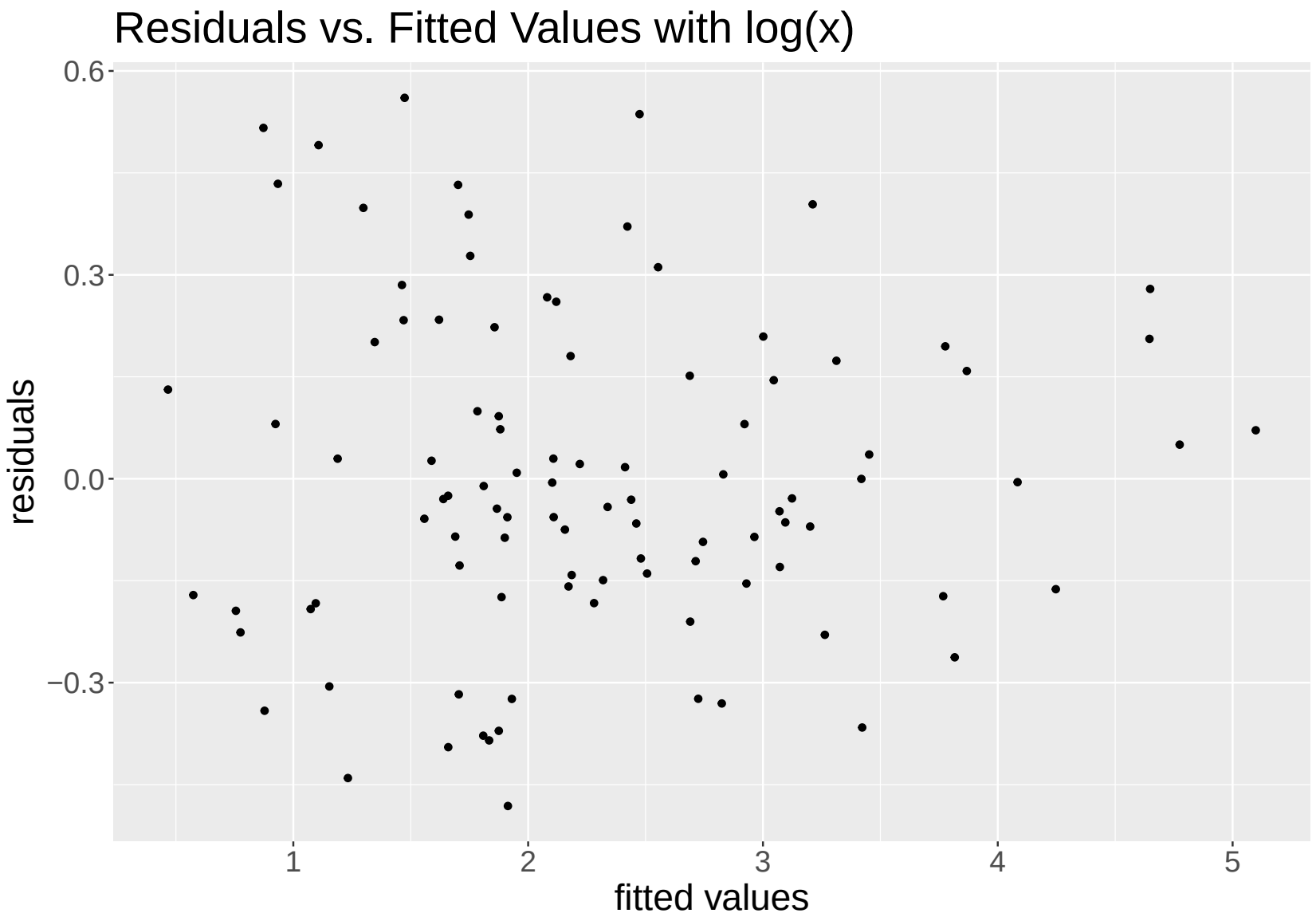
Log transformation of x

Now here is a (simulated) example where log-transforming the x-variable works really well:



Log transformation of x

And the residual plot (things look good here):



Log transformation of x

And the improvement in R^2:

Untransformed model:

```
model_x <- lm(y ~ x, data=df)
summary(model_x)$r.squared
```

```
## [1] 0.6720151
```

Transformed model:

```
model_logx <- lm(y ~ log(x), data=df)
summary(model_logx)$r.squared
```

```
## [1] 0.9435078
```

Log transformation of x

Important note: this is still LINEAR regression!

This is true even though we are modeling a curved relationship between y and x. How??

It is because we are modeling a linear relationship between y and $\log(x)$.

Also note that it is a convention among statisticians to use "log" to refer to the natural log or "ln." We will refer to the base-10 log, if ever used, as "$\log_{10}$."

Two Questions:

- What is the interpretation of the slope coefficient now?
- What are the benefits of doing the transformation, in terms of statistical inference?

Log transformation of x

Coefficient interpretation

```
##
## Call:
## lm(formula = y ~ log(x), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48146 -0.15947 -0.02937  0.16230  0.56042
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.51529    0.03061   49.51   <2e-16 ***
## log(x)      -0.96978    0.02397  -40.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2392 on 98 degrees of freedom
## Multiple R-squared:  0.9435, Adjusted R-squared:  0.9429
## F-statistic:  1637 on 1 and 98 DF,  p-value: < 2.2e-16
```

Log transformation of x

Coefficient interpretation

The value of -0.9698 is...

Log transformation of x

Statistical Power

Question: If we correctly fit a linear model with $\log(x)$ instead of just x, do we benefit in terms of statistical power?

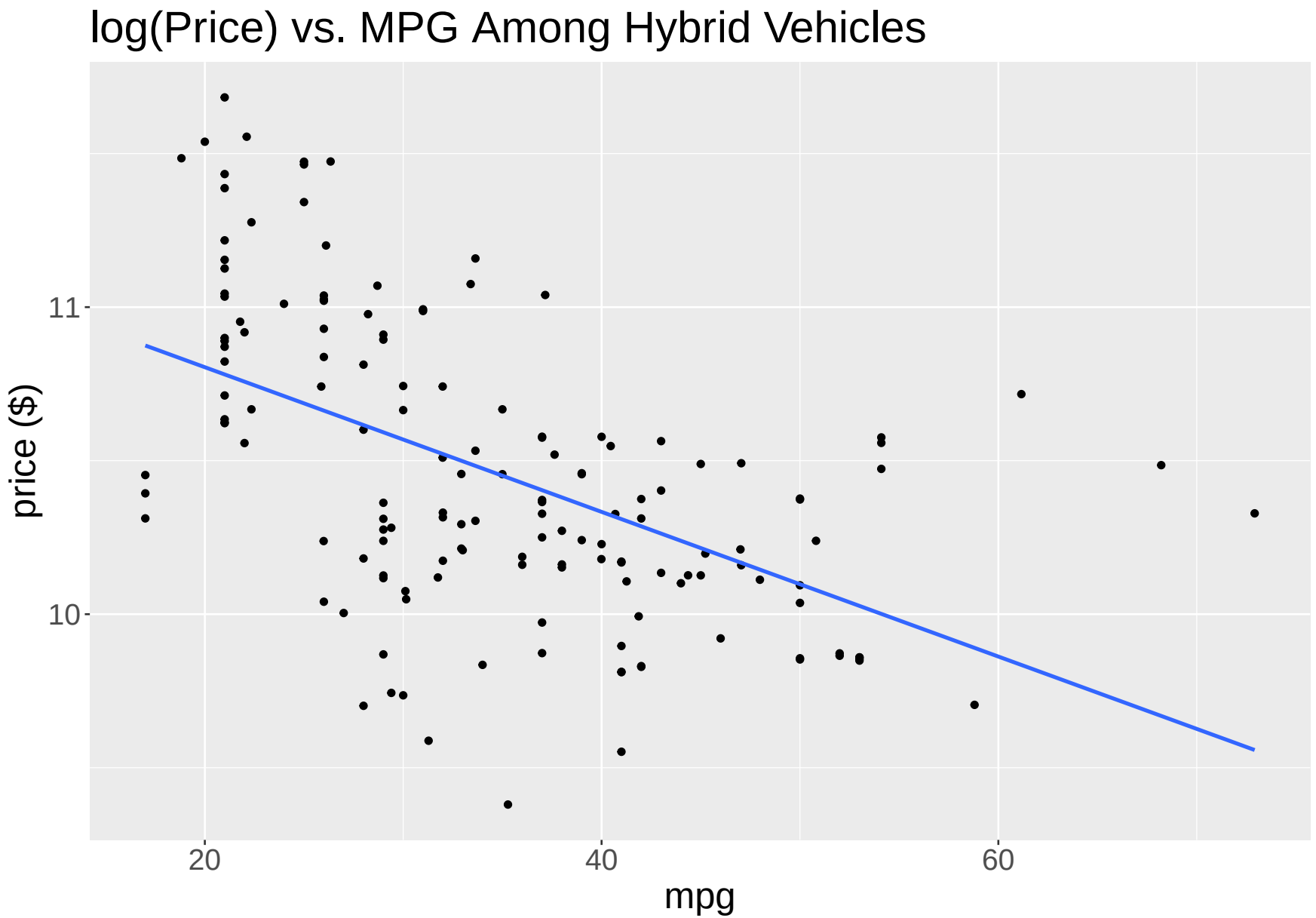
Your Turn (#1?): Simulation of power

- Simulate 100 values of $x \sim \text{Unif}(0, 3)$
- Generate values of $y = \log(x) + \epsilon$ where $\epsilon_i \sim N(\mu = 0, \sigma = 3)$
- Run these two linear models:
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$
  - $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot \log(x)$
- Check whether $H_0 \colon \beta_1 = 0$ is rejected in each case
- Do this repeatedly to get estimates of power for each model

Comment briefly on what you observe.

Back to the Hybrid Vehicle Dataset

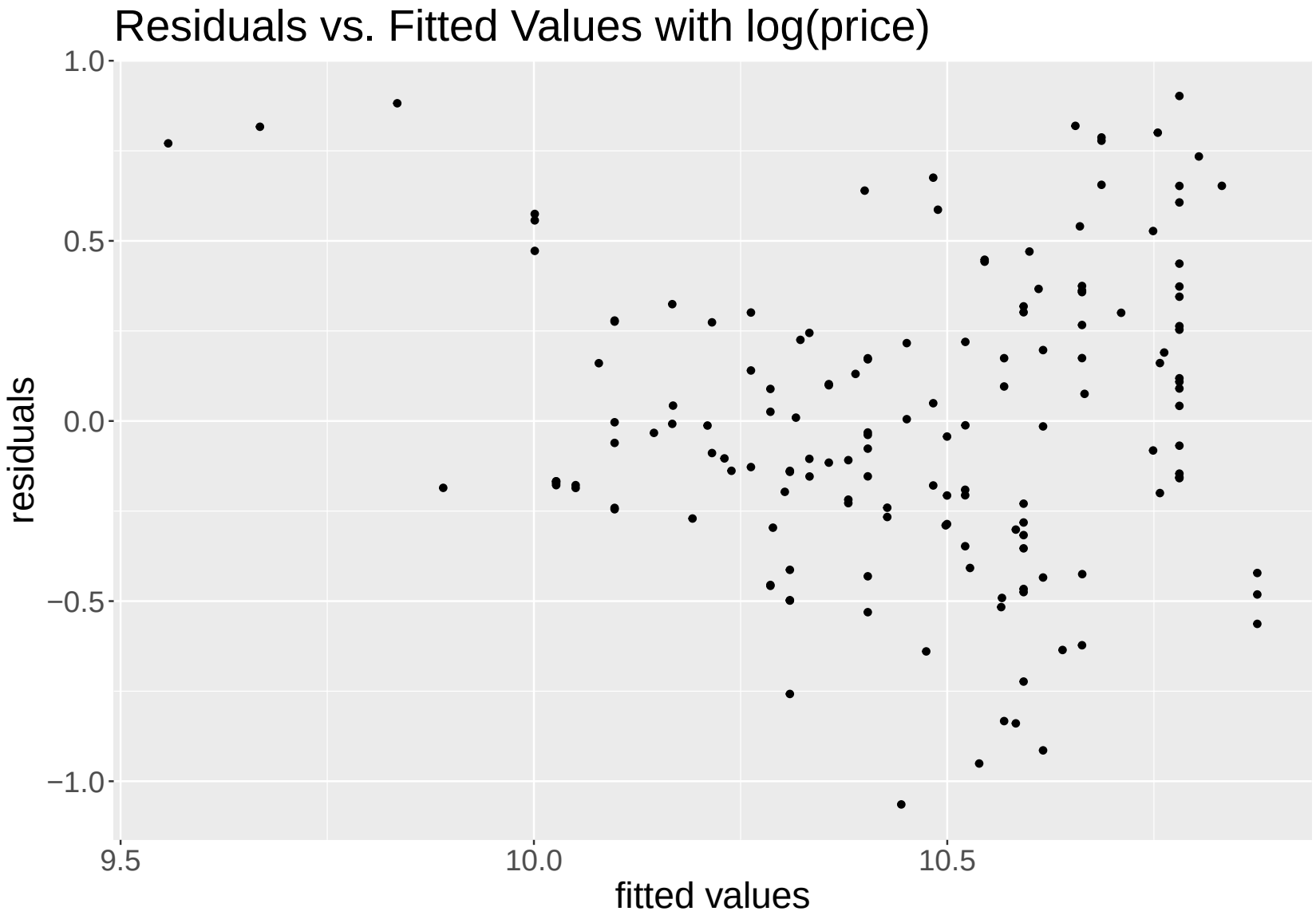
Log transformation of price



Back to the Hybrid Vehicle Dataset

Log transformation of price

This didn't help much either:



Back to the Hybrid Vehicle Dataset

Log transformation of price

And here are the R^2 values:

Untransformed model (same as before):

```
summary(model1)$r.squared
```

```
## [1] 0.2828393
```

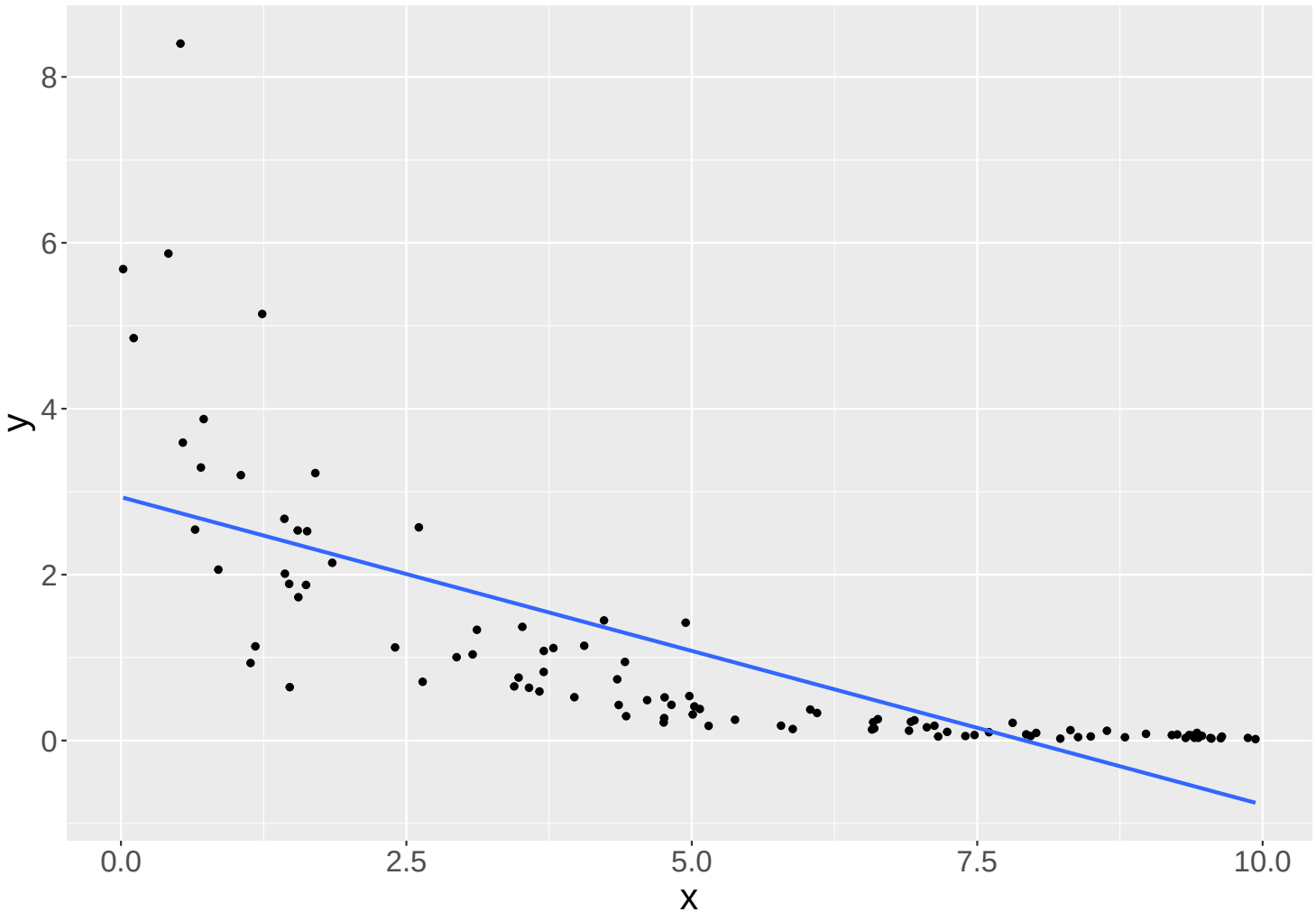
log(price) model:

```
model3 <- lm(log(price) ~ mpg, data=hybrid)
summary(model3)$r.squared
```

```
## [1] 0.2839967
```

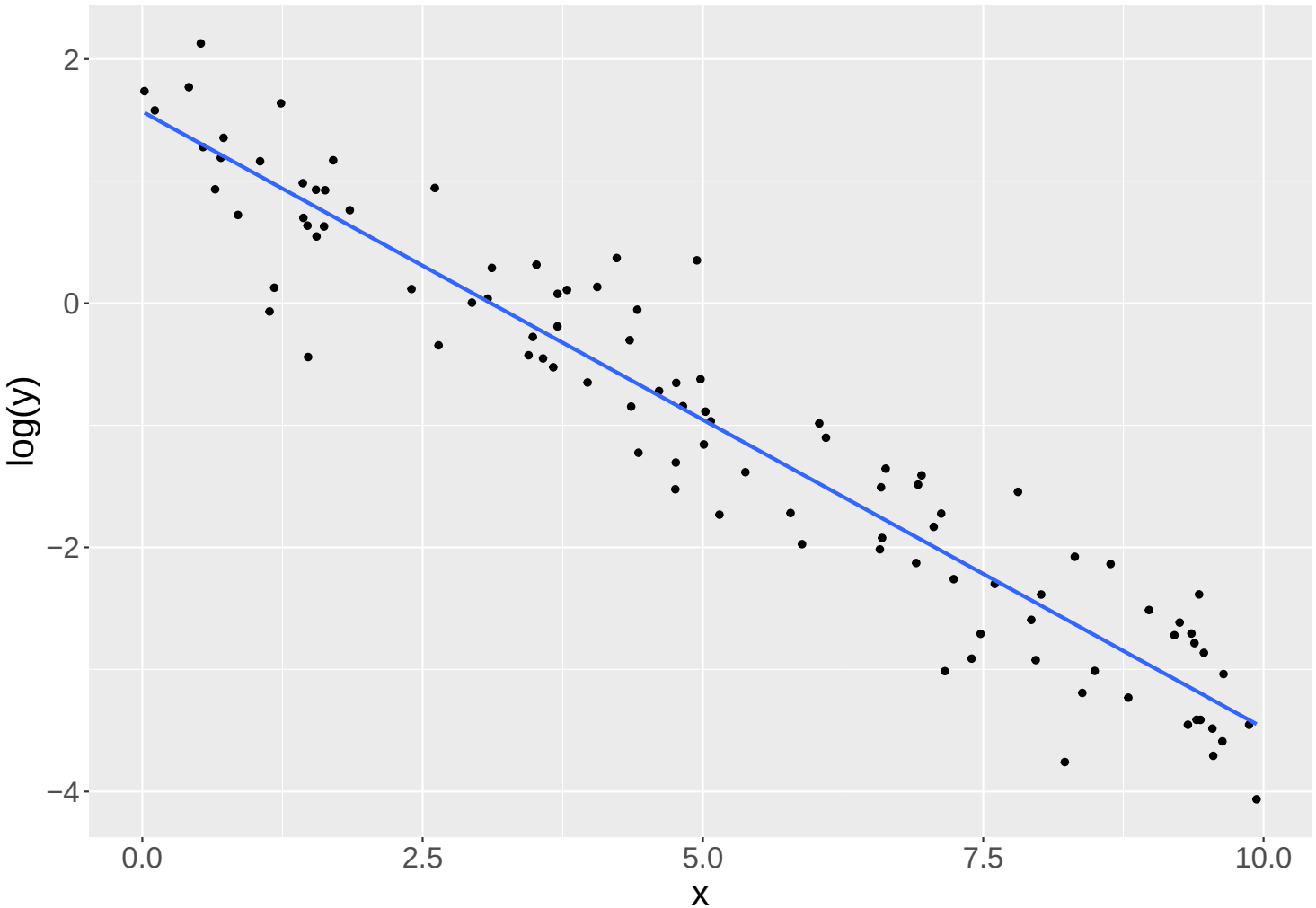
Log transformation of y

Now here is a (simulated) example where log-transforming the y-variable works really well:



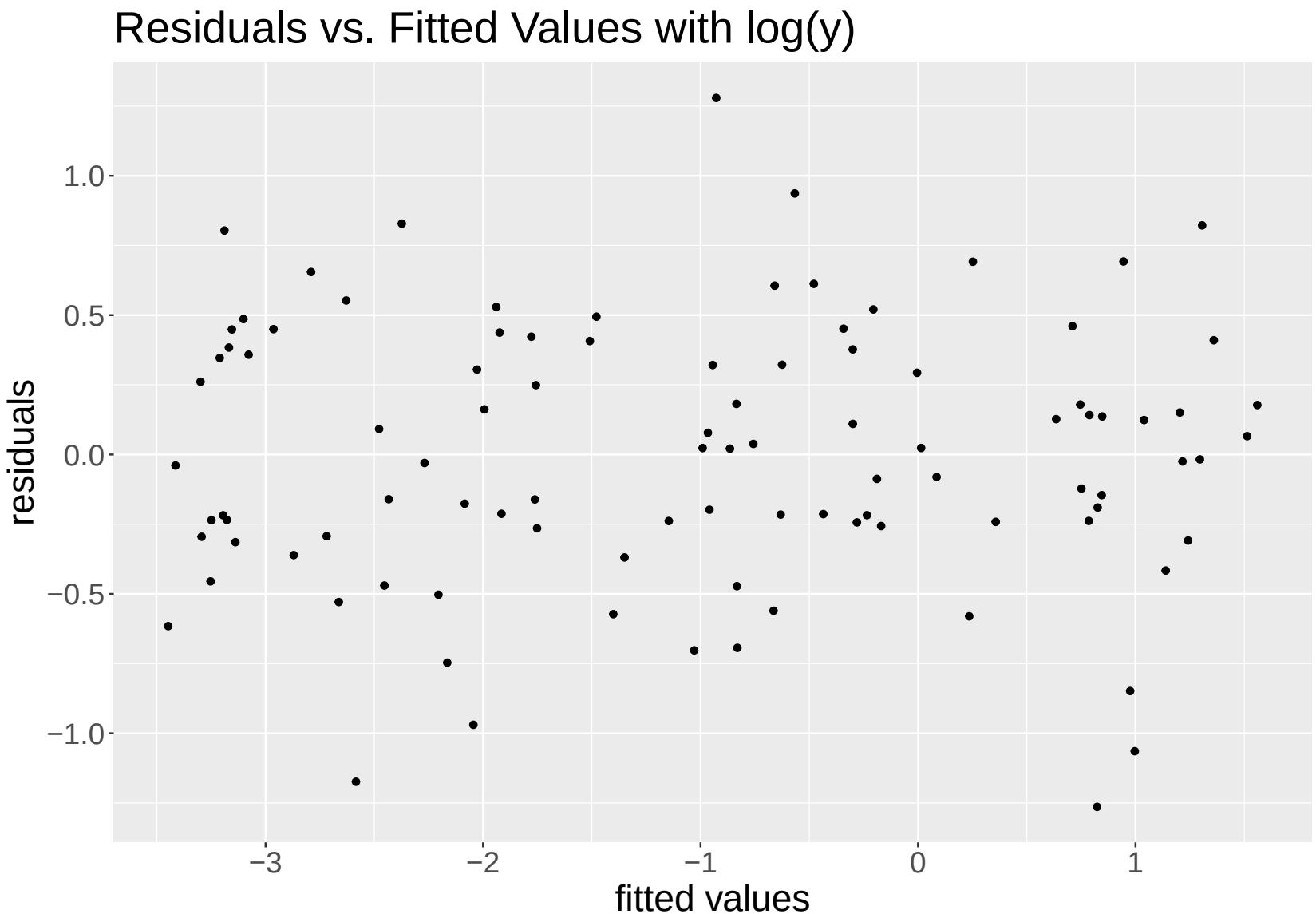
Log transformation of y

Now here is a (simulated) example where log-transforming the y-variable works really well:



Log transformation of y

And the residual plot (things look good here):



Log transformation of y

And, the improvement in R^2:

Untransformed model:

```
model_y <- lm(y ~ x, data=df)
summary(model_y)$r.squared
```

```
## [1] 0.552399
```

Transformed model:

```
model_logy <- lm(log(y) ~ x, data=df)
summary(model_logy)$r.squared
```

```
## [1] 0.9106997
```

Log transformation of y

Coefficient interpretation

```
##
## Call:
## lm(formula = log(y) ~ x, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26401 -0.25859  0.00181  0.36290  1.27910
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.56970    0.09521   16.49   <2e-16 ***
## x           -0.50490    0.01597  -31.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4779 on 98 degrees of freedom
## Multiple R-squared:  0.9107, Adjusted R-squared:  0.9098
## F-statistic: 999.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

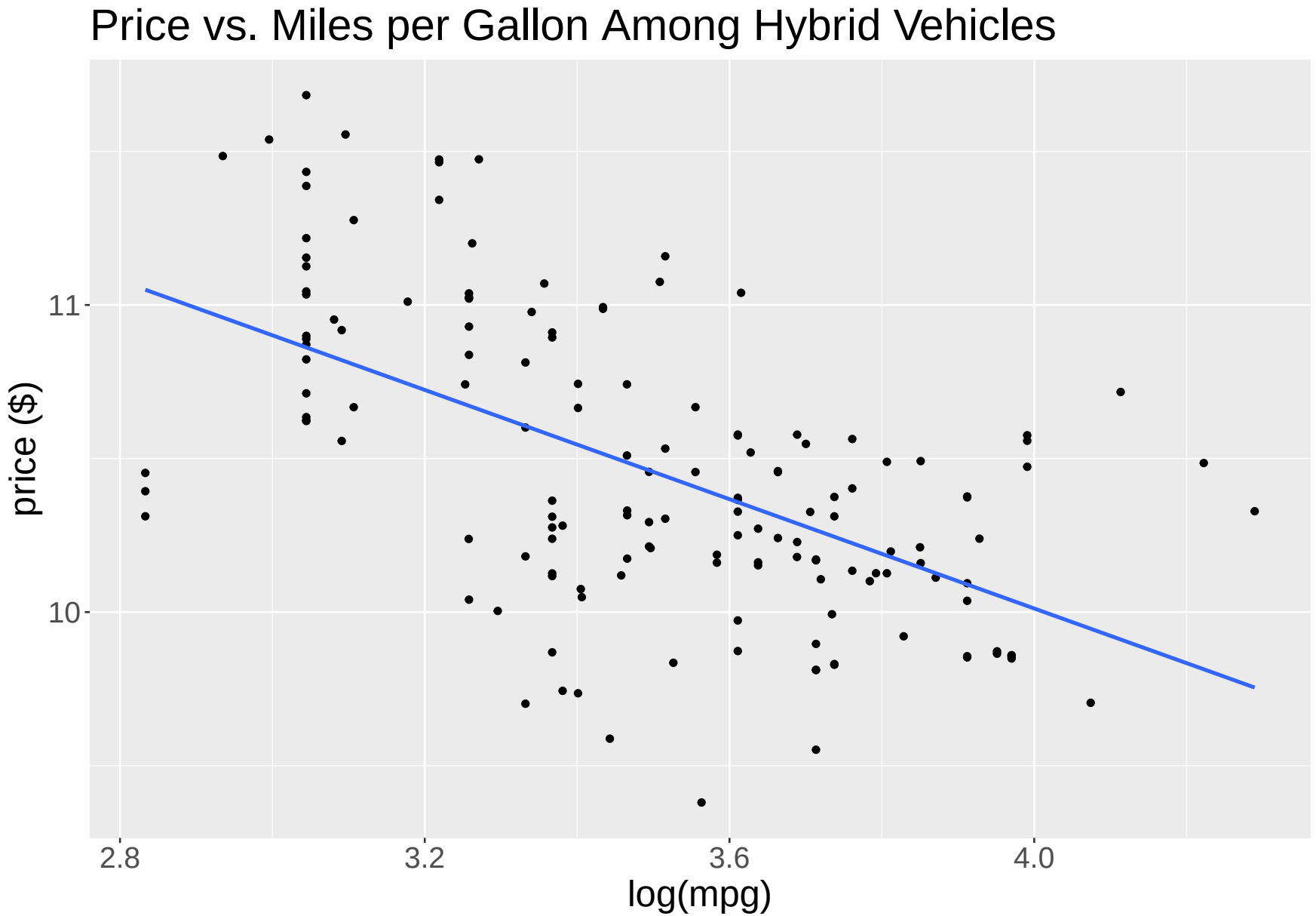
Log transformation of y

Coefficient interpretation

The value of -0.5049 is...

Back to the Hybrid Vehicle Dataset Again

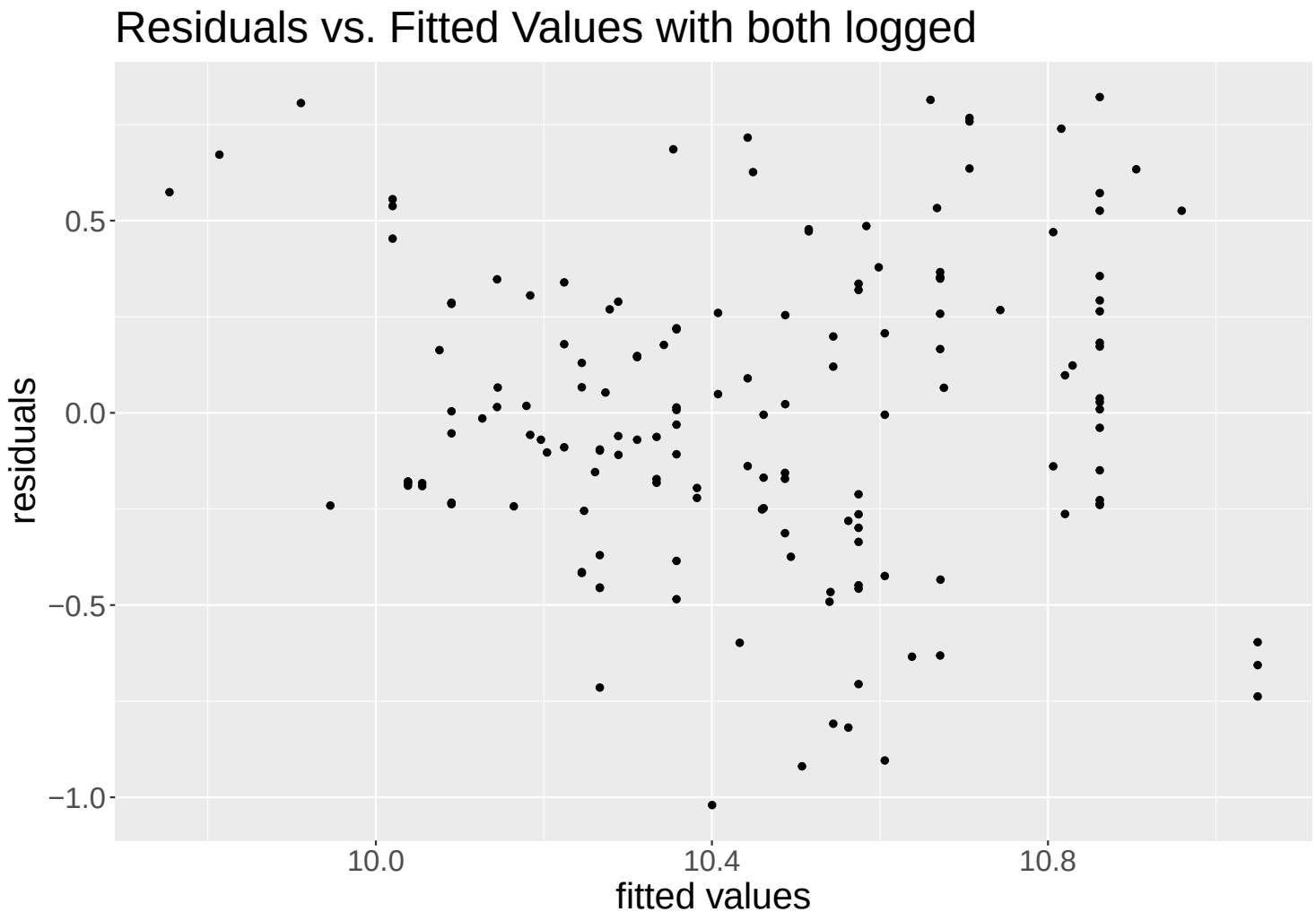
Log transformation of both



Back to the Hybrid Vehicle Dataset Again

Log transformation of both

This is actually somewhat ok!



Back to the Hybrid Vehicle Dataset Again

All the R^2 values

Untransformed model

0.2828393

log(MPG) model

0.3326834

log(price) model

0.2839967

double-log model

0.331957

So the double-log model doesn't actually have the best R^2 value, but I like its residual plot better than that of the log(MPG) model...

Back to the Hybrid Vehicle Dataset Again

Log transformation of both

```
##
## Call:
## lm(formula = log(price) ~ log(mpg), data = hybrid)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02017 -0.23954 -0.00494  0.26923  0.82136
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.5698     0.3609  37.596  < 2e-16 ***
## log(mpg)     -0.8895     0.1027  -8.662 6.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3978 on 151 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3275
## F-statistic: 75.03 on 1 and 151 DF,  p-value: 6.582e-15
```

Back to the Hybrid Vehicle Dataset Again

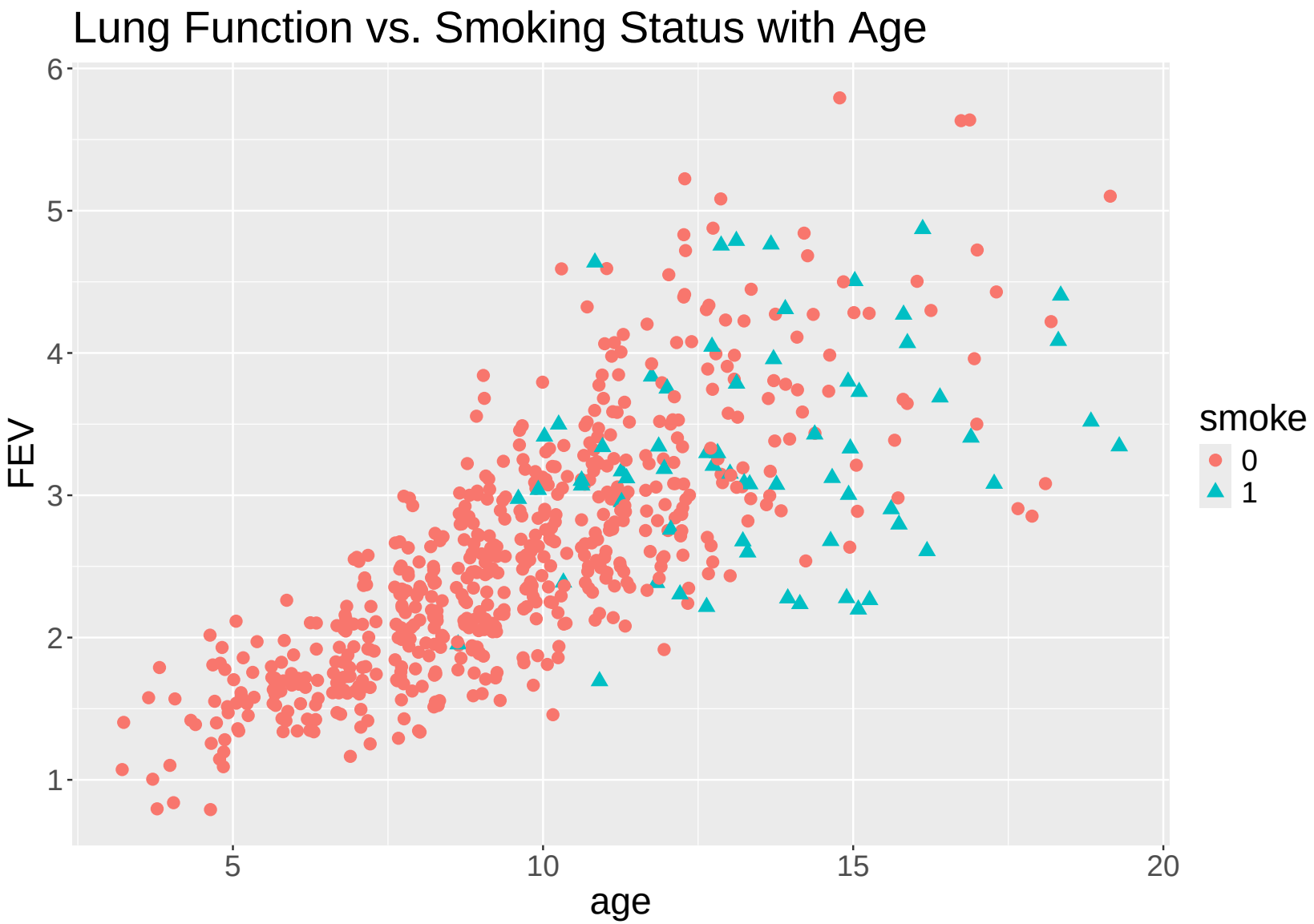
Log transformation of both

Coefficient interpretation

The value of -0.8895 is. . .

Centering and Scaling

Recall the FEV dataset from previous lectures:



Centering and Scaling

And the adjusted (no interaction) model:

```
##
## Call:
## lm(formula = fev ~ smoke + age, data = FEV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6653 -0.3564 -0.0508  0.3494  2.0894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.367373   0.081436   4.511 7.65e-06 ***
## smoke       -0.208995   0.080745  -2.588  0.00986 **
## age          0.230605   0.008184  28.176  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5651 on 651 degrees of freedom
## Multiple R-squared:  0.5766, Adjusted R-squared:  0.5753
## F-statistic: 443.3 on 2 and 651 DF,  p-value: < 2.2e-16
```

Centering and Scaling

From this output, in one sense it appears that smoking status and age have approximately the same impact on FEV, since their coefficient values are approximately of the same magnitude.

But, is this really the case?

Also, what happens to coefficient interpretations when we center and scale?

Centering and Scaling

First let's center the variables

```
FEV$age_c <- (FEV$age - mean(FEV$age))
```

Now, for a binary variable, a typical thing to do is to set:

- 0 $\rightarrow$ -0.5
- 1 $\rightarrow$ 0.5

Strictly speaking, this is not centering the variable (since its mean is not necessarily exactly 0.5). But, let's see what happens when we do this:

```
FEV$smoke_c <- ifelse(FEV$smoke==0, -0.5, 0.5)
```

Centering and Scaling

First let's center the variables

```
model_c <- lm(fev ~ smoke_c + age_c, data=FEV)
summary(model_c)
```

```
##
## Call:
## lm(formula = fev ~ smoke_c + age_c, data = FEV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6653 -0.3564 -0.0508  0.3494  2.0894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.553054   0.039174  65.173  < 2e-16 ***
## smoke_c     -0.208995   0.080745  -2.588  0.00986 **
## age_c        0.230605   0.008184  28.176  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5651 on 651 degrees of freedom
## Multiple R-squared:  0.5766,  Adjusted R-squared:  0.5753
```

Centering and Scaling

First let's center the variables

What changed, and why?

Centering and Scaling

Now let's also scale age

```
FEV$age_s <- (FEV$age - mean(FEV$age))/(2*sd(FEV$age))
```

Wait why divide by 2sd?

Centering and Scaling

Now let's also scale age

```
model_s <- lm(fev ~ smoke_c + age_s, data=FEV)
summary(model_s)
```

```
##
## Call:
## lm(formula = fev ~ smoke_c + age_s, data = FEV)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6653 -0.3564 -0.0508  0.3494  2.0894
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.55305    0.03917  65.173  < 2e-16 ***
## smoke_c     -0.20899    0.08075  -2.588  0.00986 **
## age_s        1.36238    0.04835  28.176  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5651 on 651 degrees of freedom
## Multiple R-squared:  0.5766,  Adjusted R-squared:  0.5753
```

Centering and Scaling

When we center and scale the covariates, a couple of things happen:

1) Coefficients become more directly comparable to each other
2) The intercept term is now the estimated value of the outcome variable at the average value of the predictor variables (as opposed to when the predictor variables equal 0)

Your Turn #2

What happens in the interaction model??

Centering and Scaling

Your Turn #2

- Load the FEV dataframe into R
- Run the interaction model with age and smoking status (using the original untransformed values)
- Then center both variables and re-run the interaction model.
  - Briefly comment on what has changed, and give new interpretations of each regression coefficient
- Finally, scale the age variable and re-run the interaction model
  - Again briefly comment on what has changed, and give new interpretations of each regression coefficient

Recap and Looking Ahead

Recap

- Transformations (e.g. log, and others not covered here) can fix violations to the linearity condition
- Transformations (e.g. centering and scaling) can also be helpful with interpretations of regression coefficients

Looking Ahead

Logistic Regression

Today's Daily Check

Both Your Turns.