STATS 2107
Statistical Modelling and Inference II
Solutions
Workshop 10:
Transformations in MLR

Matt Ryan

Semester 2 2022

Contents

The MLR approach

The data

We will look at the `trees` dataset built into R. This data set provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees

```
tribble(
  ~Variable, ~Description,
  "Girth", "Tree diameter in inches",
  "Height", "Tree height in feet",
  "Volume", "Tree volume of timber in cubic feet"
) %>%
  knitr::kable()
```

Variable	Description
Girth	Tree diameter in inches
Height	Tree height in feet
Volume	Tree volume of timber in cubic feet

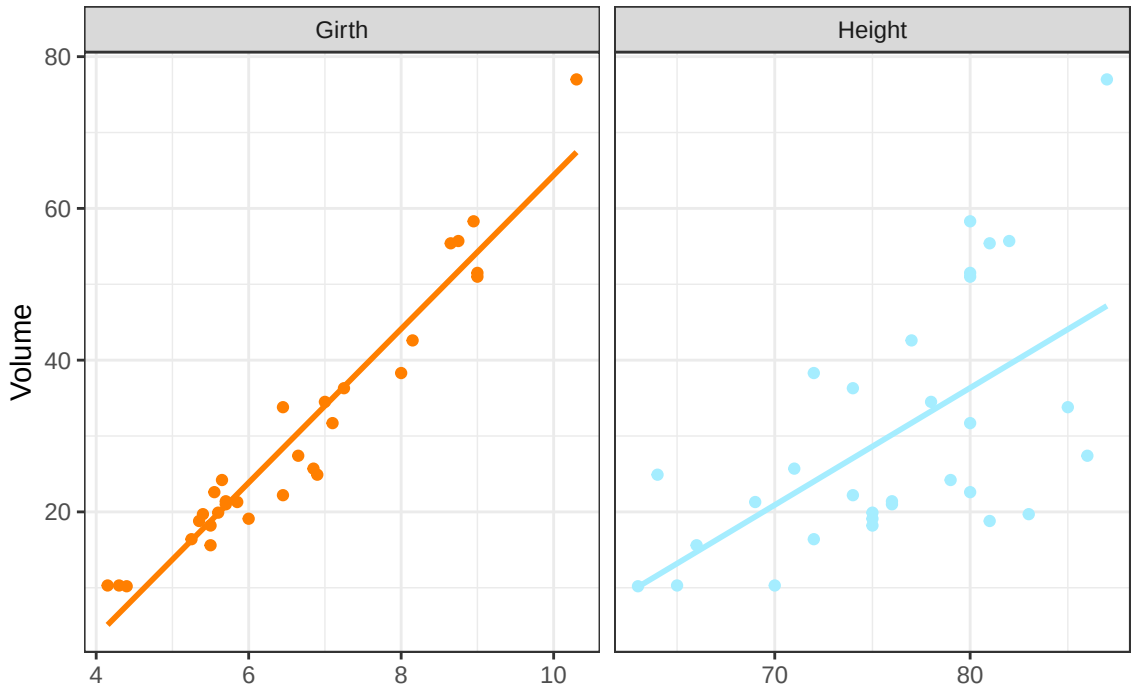
Load the data

```
data("trees")
trees <- trees %>%
  as_tibble() %>%
  mutate(Girth = Girth/2) # Let's look at radius, not diameter
head(trees)
```

```
## # A tibble: 6 x 3
##   Girth Height Volume
##   <dbl>  <dbl>  <dbl>
## 1  4.15     70   10.3
## 2  4.3      65   10.3
## 3  4.4      63   10.2
## 4  5.25     72   16.4
## 5  5.35     81   18.8
## 6  5.4      83   19.7
```

Look at the data

```
trees %>%
  pivot_longer(-Volume) %>%
  ggplot(aes(x = value, y = Volume, colour = name)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x, se = FALSE) +
  scale_colour_manual(values = cols) + # This code won't work for you
  facet_wrap( ~ name, scales = "free_x", nrow = 1) +
  theme_bw() +
  labs(x = NULL) +
  theme(legend.position = "none")
```



Let's fit a linear model

To predict `Volume`, let's fit the following model

```
trees_lm1 <- lm(Volume ~ Height + Girth, data = trees)
summary(trees_lm1)
```

```
##
## Call:
## lm(formula = Volume ~ Height + Girth, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## Girth         9.4163     0.5285  17.816  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

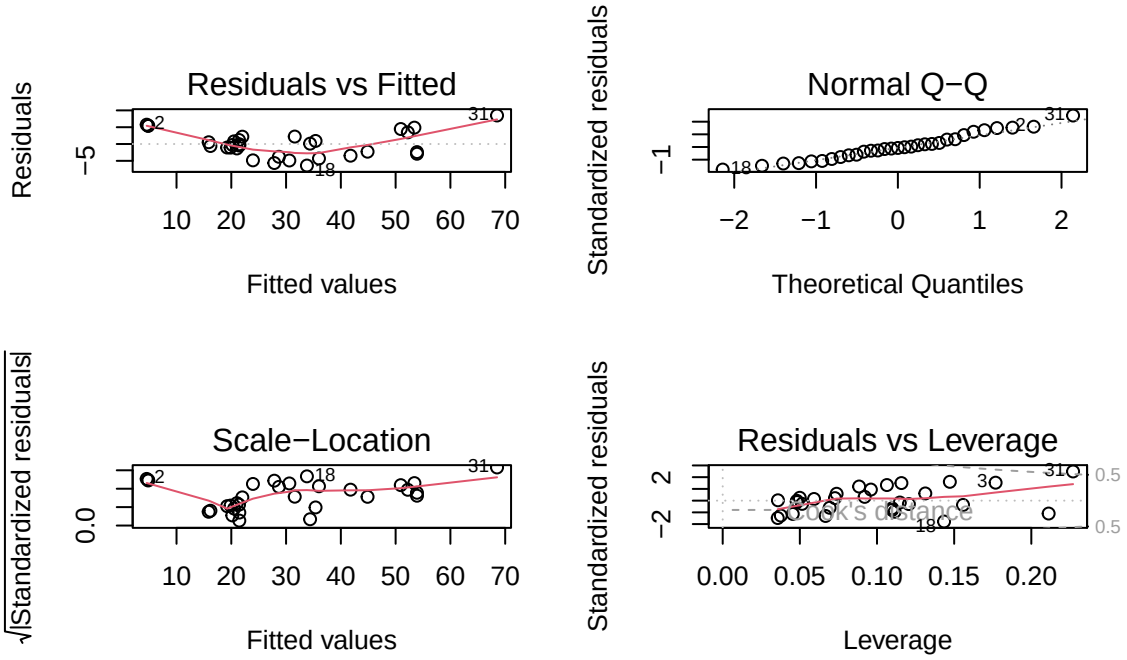
Your turn

What to do

1. Does this model meet the assumptions of linear regression?

Solutions:
First, let's get the plots:

```
par(mfrow = c(2, 2))
plot(trees_lm1)
```

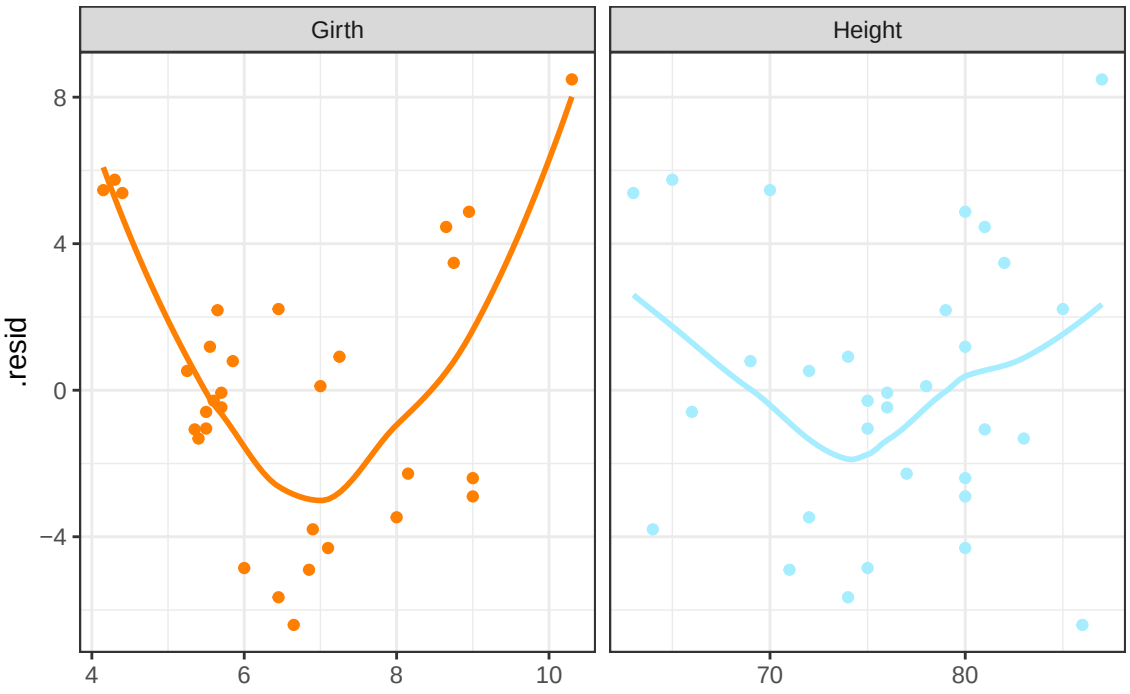


Solutions:
This is quite clearly not linear.

2. Look at the residuals versus each predictor. Does this tell you anything?

Solutions:

```
broom::augment(trees_lm1) %>%
  select(Height, Girth, .resid) %>%
  pivot_longer(-.resid) %>%
  ggplot(aes(x = value, y = .resid, colour = name)) +
  geom_point() +
  geom_smooth(method = loess, formula = y ~ x, se = FALSE) +
  scale_colour_manual(values = cols) + # This code won't work for you
  facet_wrap( ~ name, scales = "free_x", nrow = 1) +
  theme_bw() +
  labs(x = NULL) +
  theme(legend.position = "none")
```



Solutions:
Looking at these, we get clear non-linearity in both predictors.

A more informed approach

What is a tree?

Question: What is a good way to estimate the volume of a tree?

Solutions:
Trees are fairly cylindrical, so the volume of a cylinder is probably smart

The volume of a cylinder

Let's estimate the volume of timber we will get from a tree as:

$$V = 2\pi h r^2$$

where

- V is the volume of the tree
- r is the radius of the tree
- h is the height of the tree

How do we linearise this?

Linearising the volume

$$\log(V) = \log(2\pi) + \log(h) + 2\log(r)$$

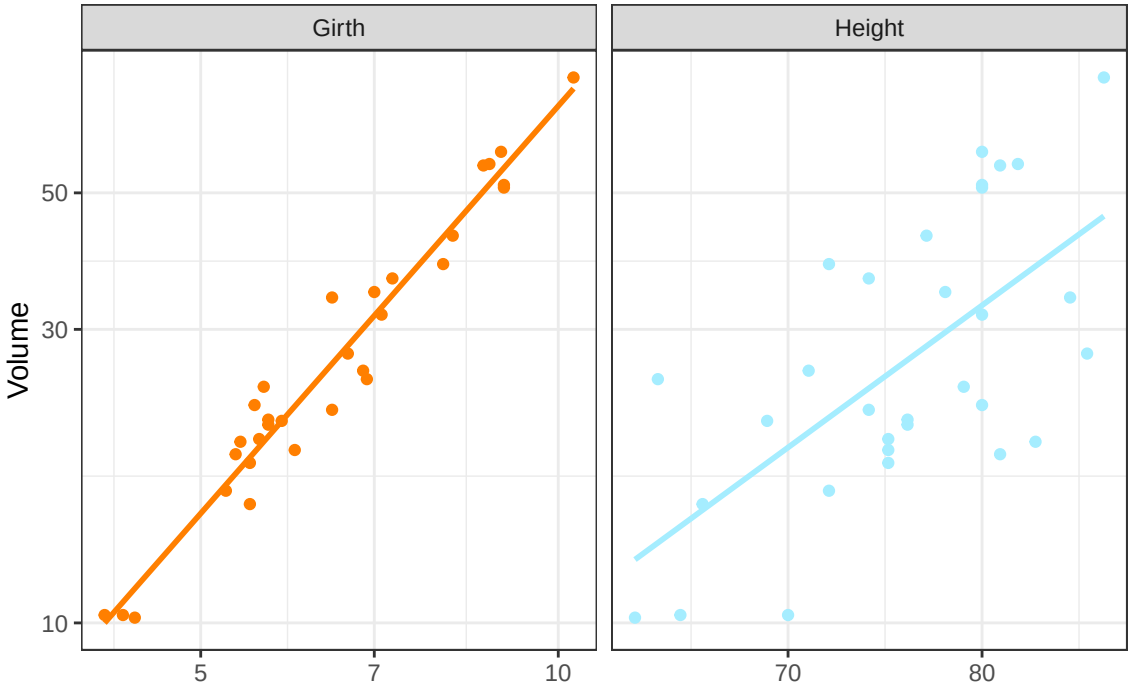
So let's consider the linear regression

$$\log(V) = \beta_0 + \beta_1 \log(h) + \beta_2 \log(r) + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$.

Does the picture look right?

```
trees %>%
  pivot_longer(-Volume) %>%
  ggplot(aes(x = value, y = Volume, colour = name)) +
  geom_point() +
  geom_smooth(method = lm, formula = y ~ x, se = FALSE) +
  scale_colour_manual(values = cols) + # This code won't work for you
  facet_wrap( ~ name, scales = "free_x", nrow = 1) +
  scale_x_log10()+
  scale_y_log10()+
  theme_bw() +
  labs(x = NULL) +
  theme(legend.position = "none")
```



DISCLAIMER: This is actually $\log_{10}$ scale.

Fit the model

```
trees_lm2 <- lm(log(Volume) ~ log(Height) + log(Girth), data = trees)
summary(trees_lm2)
```

```
##
## Call:
## lm(formula = log(Volume) ~ log(Height) + log(Girth), data = trees)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168561 -0.048488  0.002431  0.063637  0.129223
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.25735    0.81926  -6.417 6.00e-07 ***
## log(Height)  1.11712    0.20444   5.464 7.81e-06 ***
## log(Girth)   1.98265    0.07501  26.432  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08139 on 28 degrees of freedom
## Multiple R-squared:  0.9777, Adjusted R-squared:  0.9761
## F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

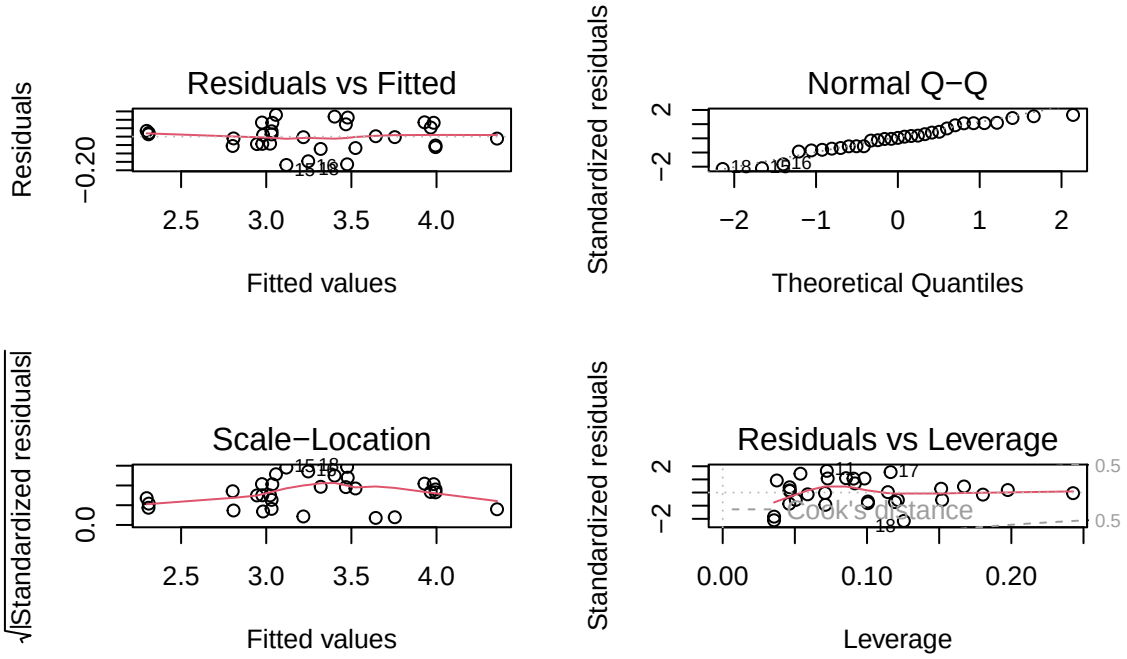
Your turn

What to do

1. Does this model meet the assumptions of linear regression?

Solutions:
First, let's get the plots:

```
par(mfrow = c(2, 2))
plot(trees_lm2)
```



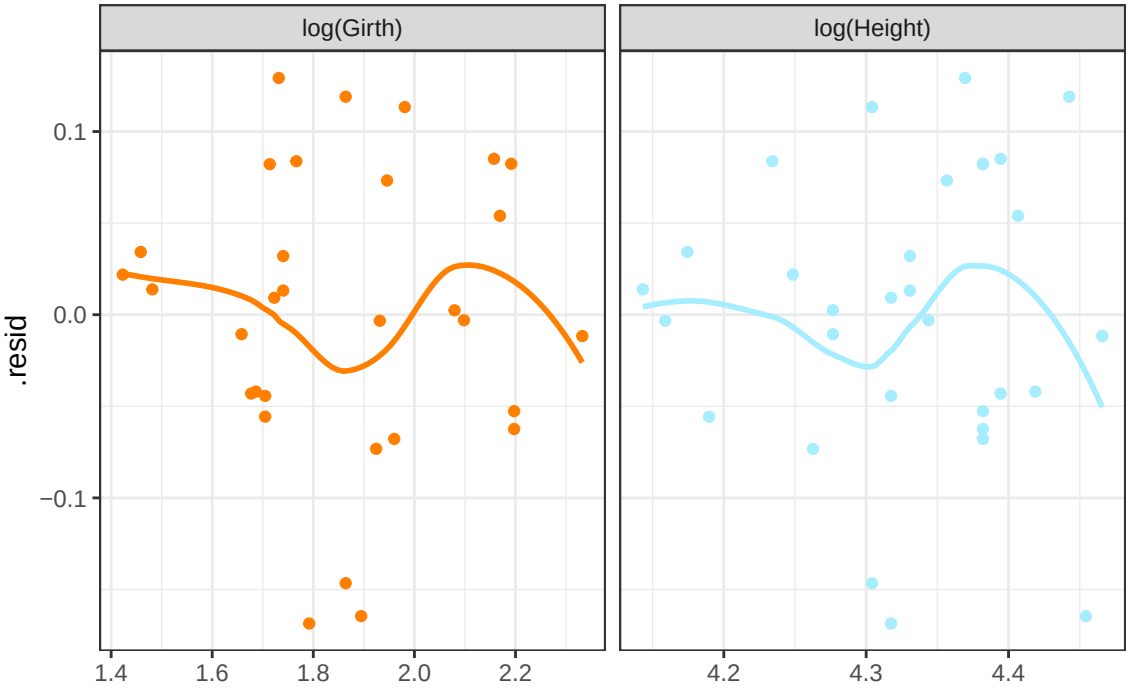
Solutions:

Looking much better. Maybe some lack of constant variance, but this could be explained by few data points at the end

2. Look at the residuals versus each predictor. Does this tell you anything?

Solutions:

```
trees %>%
  add_column(.resid = residuals(trees_lm2)) %>%
  mutate(`log(Girth)` = log(Girth),
         `log(Height)` = log(Height)) %>%
  select(`log(Height)`, `log(Girth)`, .resid) %>%
  pivot_longer(-.resid) %>%
  ggplot(aes(x = value, y = .resid, colour = name)) +
  geom_point() +
  geom_smooth(method = loess, formula = y ~ x, se = FALSE) +
  scale_colour_manual(values = cols) + # This code won't work for you
  facet_wrap( ~ name, scales = "free_x", nrow = 1) +
  theme_bw() +
  labs(x = NULL) +
  theme(legend.position = "none")
```



Solutions:
Better, but not perfect. This could again be explained by the lack of data in the endpoints.

3. To see if our model is a smart choice, test the hypotheses (at the 5% level):

$$H_0 : \beta_1 = 1 \quad \text{vs} \quad H_a : \beta_1 \neq 1\,,$$

and

$$H_0 : \beta_2 = 2 \quad \text{vs} \quad H_a : \beta_2 \neq 2\,,$$

Solutions:
To test the first hypothesis, we look at the $\log(Height)$ row of the summary table. Our test statistic is:

$$T = \frac{\hat{\beta}_1 - 1}{SE(\hat{\beta}_1)} \sim t_{28}$$

under the null hypothesis. The observed value is:

$$T = \frac{1.11712 - 1}{0.20444} \approx 0.5729$$

Since this is less than $t_{28,0.025} = 2.048407$, we have insufficient evidence to reject the null hypothesis at the 5% level.

To test the first hypothesis, we look at the $\log(Girth)$ row of the summary table. Our test statistic is:

$$T = \frac{\hat{\beta}_2 - 2}{SE(\hat{\beta}_2)} \sim t_{28}$$

under the null hypothesis. The observed value is:

$$T = \frac{1.98265 - 2}{0.07501} \approx -0.2313025$$

Since this is less than $t_{28,0.025} = 2.048407$ in absolute value, we have insufficient evidence to reject the null hypothesis at the 5% level.

Our conclusion from this is the we have insufficient evidence to suggest that our model is different to that of the model of a cylinder.

Let's talk about intervals

Think about prediction

Consider data $Y_1, Y_2, \ldots, Y_n$ and the multiple linear regression model (in matrix form)

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}\,.$$

Then $Y_i \sim N(\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2)$ independently for each $i = 1, 2, \ldots, n$.

Think about prediction

Consider a new independent observation Y_0 with predictor $\boldsymbol{x}_0$, then $Y_0 \sim N(\boldsymbol{x}_0^T\boldsymbol{\beta}, \sigma^2)$.

Our best guess of Y_0 is our *predicted value* $\hat{Y}_0 = \boldsymbol{x}_0^T\hat{\boldsymbol{\beta}}$, then $\hat{Y}_0 \sim N(\boldsymbol{x}_0^T\boldsymbol{\beta}, \sigma^2\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0)$.

What is a confidence interval?

A $(1-\alpha)\times 100\%$ **confidence interval** for our new data point Y_0 is an interval describing how sure we are about the *mean value of* Y_0.

That is, it is an interval about $\mathrm{E}[Y_0] = \boldsymbol{x}_0^T\boldsymbol{\beta}$.

How to construct the confidence interval

We are trying to get an idea about $\mathrm{E}[Y_0]$, so our best guess at this value is $\hat{Y_0}$.

So how far off is $\hat{Y_0}$ from $\mathrm{E}[Y_0]$? How much to we expect $\hat{Y_0}$ to vary from $\mathrm{E}[Y_0]$?

How to construct the confidence interval

$$\hat{Y_0} - \mathrm{E}[Y_0] \sim N(0, \sigma^2 \boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0)$$

hence our $(1-\alpha) \times 100\%$ CI is

$$\hat{Y_0} \pm z_{\alpha/2} \sigma \sqrt{\boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0}$$

if σ^2 is known (replace σ^2 be s_e^2 and $z_{\alpha/2}$ by the appropriate t-critical value if σ^2 unknown.)

Can we back transform this?

Suppose $Y_0 = f(W_0)$ where f is increasing monotonic. When we get a confidence interval for $\mathrm{E}[Y_0]$, can we say anything about $\mathrm{E}[W_0]$?

Can we back transform this?

In general, no. This is because our CI is

$$L = \hat{Y_0} - z_{\alpha/2} \sigma \sqrt{\boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0} < \mathrm{E}[Y_0] < \hat{Y_0} + z_{\alpha/2} \sigma \sqrt{\boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0} = U$$

Applying f^{-1} gives

$$f^{-1}(L) < f^{-1}(\mathrm{E}[Y_0]) < f^{-1}(U)$$

and in general $f^{-1}(\mathrm{E}[Y_0]) \neq \mathrm{E}[f^{-1}(Y_0)] = \mathrm{E}[W_0]$.

What is a prediction interval?

A $(1-\alpha) \times 100\%$ **prediction interval** for our new data point Y_0 is an interval describing how sure we are about the *value of* Y_0, not it's mean!

That is, it is an interval about Y_0 itself.

Recall that Y_0 is random, with $Y_0 \sim N(\boldsymbol{x}_0^T \boldsymbol{\beta}, \sigma^2)$

How to construct the prediction interval

We are trying to get an idea about Y_0, so our best guess at this value is $\hat{Y_0}$.

So how far off is $\hat{Y_0}$ from Y_0? How much to we expect $\hat{Y_0}$ to vary from Y_0?

How to construct the prediction interval

$$\hat{Y_0} - Y_0 \sim N(0, \sigma^2 (1 + \boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0))$$

hence our $(1-\alpha) \times 100\%$ PI is

$$\hat{Y_0} \pm z_{\alpha/2} \sigma \sqrt{1 + \boldsymbol{x}_0^T (X^T X)^{-1} \boldsymbol{x}_0}$$

if σ^2 is known (replace σ^2 be s_e^2 and $z_{\alpha/2}$ by the appropriate t-critical value if σ^2 unknown.)

Can we back transform this?

Suppose $Y_0 = f(W_0)$ where f is increasing monotonic. When we get a prediction interval for Y_0, can we say anything about W_0?

Can we back transform this?

Yes we can! This is because our PI is

$$L = \hat{Y_0} - z_{\alpha/2}\sigma\sqrt{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0} < Y_0 < \hat{Y_0} + z_{\alpha/2}\sigma\sqrt{1 + \boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0} = U$$

Applying f^{-1} gives

$$f^{-1}(L) < f^{-1}(Y_0)(= W_0) < f^{-1}(U)$$

Your turn

What to do

1. Using the transformed model from before, obtain a 95% confidence interval for a cherry tree with height 80 feet, and radius of 8 inches.

Solutions:
First, let's set up the new data frame. Now, R is smart, so we can give this data on the *original* scale.

```
new_tree <- tibble(
  Girth = 8,
  Height = 80
)
new_tree
```

```
## # A tibble: 1 x 2
##   Girth Height
##   <dbl>  <dbl>
## 1     8     80
```

Solutions:
Now to get the confidence interval, let's use the `predict` function:

```
predict(trees_lm2, newdata = new_tree, interval = "confidence")
```

```
##       fit      lwr      upr
## 1 3.76072 3.719341 3.802099
```

Solutions:
So our CI on the *transformed scale* is $(3.719341, 3.802099)$.

2. Using the transformed model from before, obtain a 95% prediction interval for a cherry tree with height 80 feet, and radius of 8 inches.

Solutions:
Using the same data set up as before, we can get a PI using the `predict` function and changing the interval type.

```
predict(trees_lm2, newdata = new_tree, interval = "prediction")
```

```
##       fit     lwr     upr
## 1 3.76072 3.58895 3.93249
```

Solutions:
So our PI on the *transformed scale* is $(3.58895, 3.93249)$.

3. What can we say about the volume of the cherry tree from these intervals?

Solutions:
We cannot say anything from the confidence interval, but we can back-transform the prediction interval using the exponential as:

```
exp(predict(trees_lm2, newdata = new_tree, interval = "prediction"))
```

```
##        fit      lwr     upr
## 1 42.97935 36.19603 51.0339
```

Solutions:
Hence, we are 95% confident that the volume of a black cherry tree with a height of 80 feet and a radius of 8 inches will be between 36.19603 and 51.0339 cubic feet.