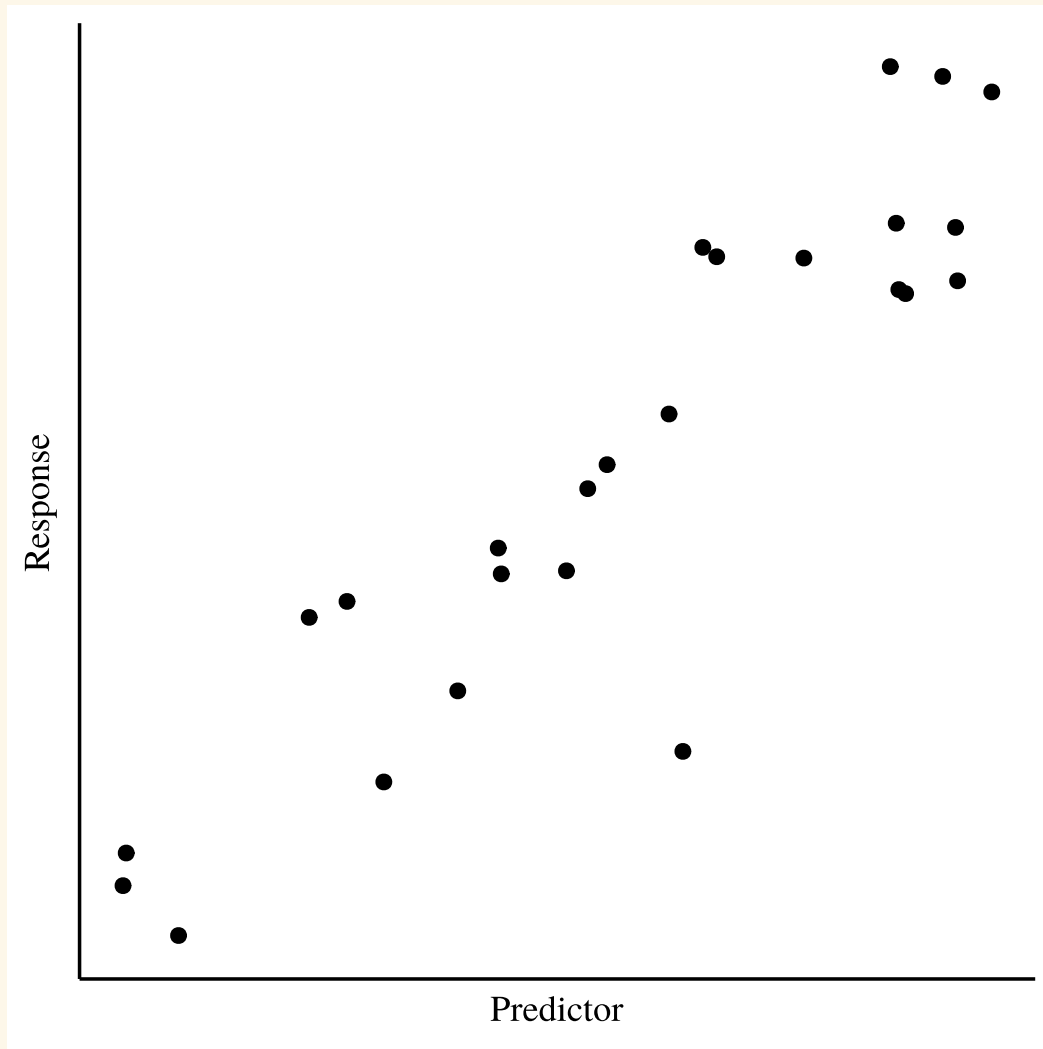


SLR: R-squared and brief ANOVA intro

Stat 230

April 13 2022

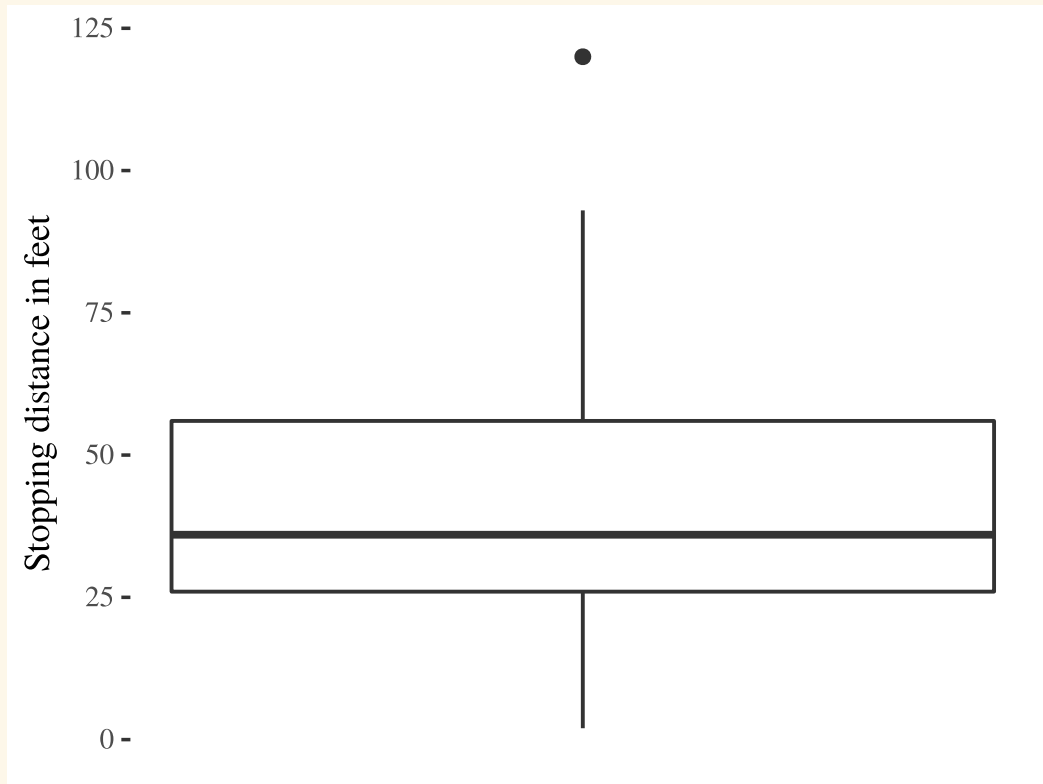
Overview



Today:

- Why include predictors in SLR?
- Percent variability explained (R-squared)
- Simple Analysis of Variance (ANOVA)

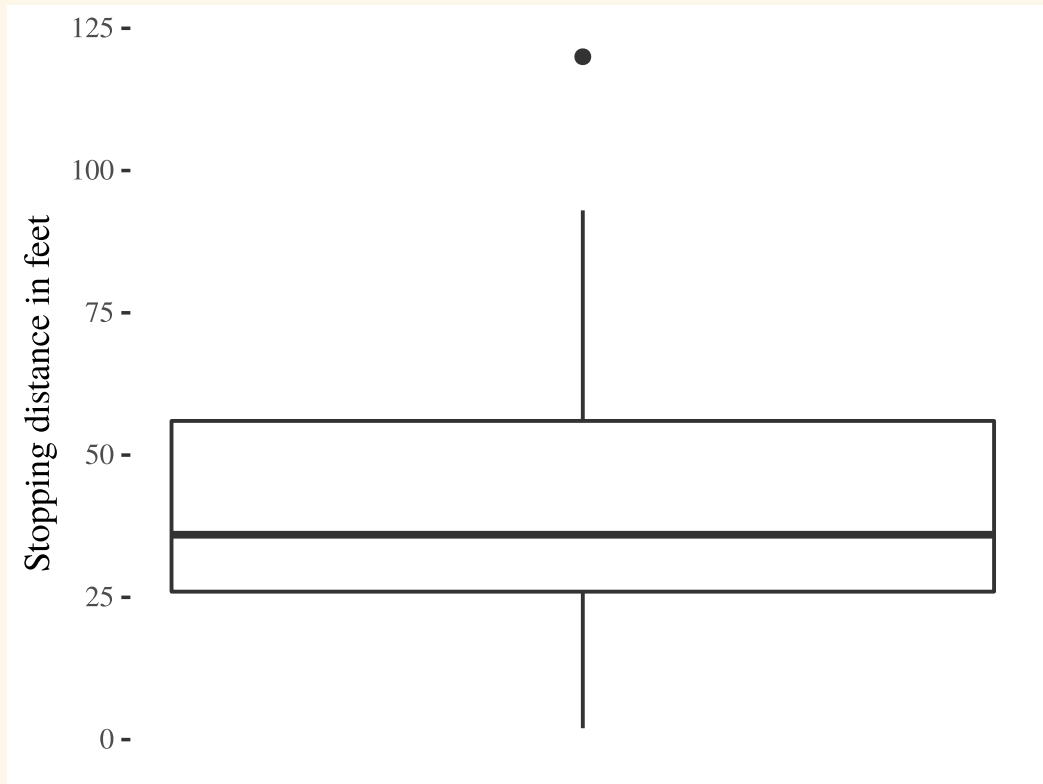
Example: Cars data



Suppose we want to guess the stopping distance of cars in 1920s but don't have the speeds.

- How much variability?

Example: Cars data



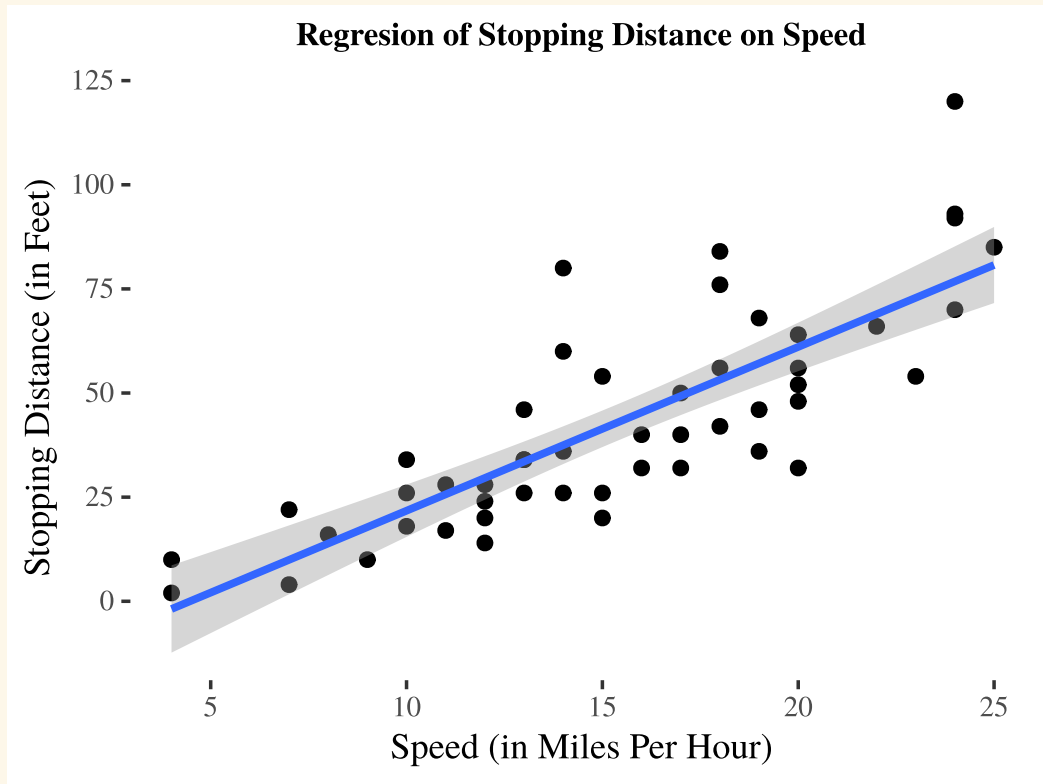
Suppose we want to guess the stopping distance of cars in 1920s but don't have the speeds.

- How much variability?

Stopping distances range from about 0 feet to 125 *feet*, with a standard deviation of about 26 *feet* and variance of about 664 *feet*²

```
sd(cars$dist) # standard deviation
[1] 25.76938
sd(cars$dist)^2 # variance
[1] 664.0608
```

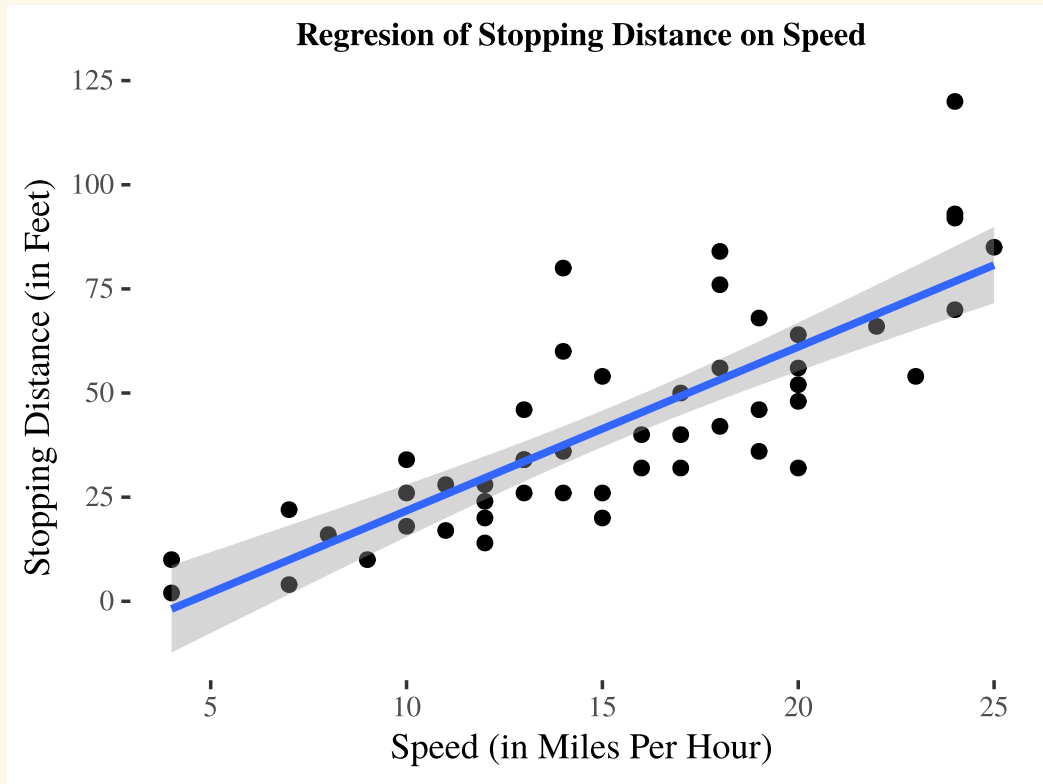
Example: Cars data



Now suppose we know the speeds and want to guess the depth.

- How much variability?

Example: Cars data



Now suppose we know the speeds and want to guess the depth.

- How much variability?

At any speed, the standard deviation of stopping distance is about 15.4 *feet* and variance is about 236.5 *feet*²

```
cars_lm <- lm(dist ~ speed, data = cars)
summary(cars_lm)$sigma
[1] 15.37959
summary(cars_lm)$sigma^2
[1] 236.5317
```

R-squared

R-squared (R^2 or coefficient of determination) measures the proportion of variability observed in the response Y which can be explained by the regression of Y on x.

```
summary(cars_lm)
```

```
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min       1Q   Median       3Q      Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed        3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,    Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

R-squared

```
summary(cars_lm)$r.squared  
[1] 0.6510794
```

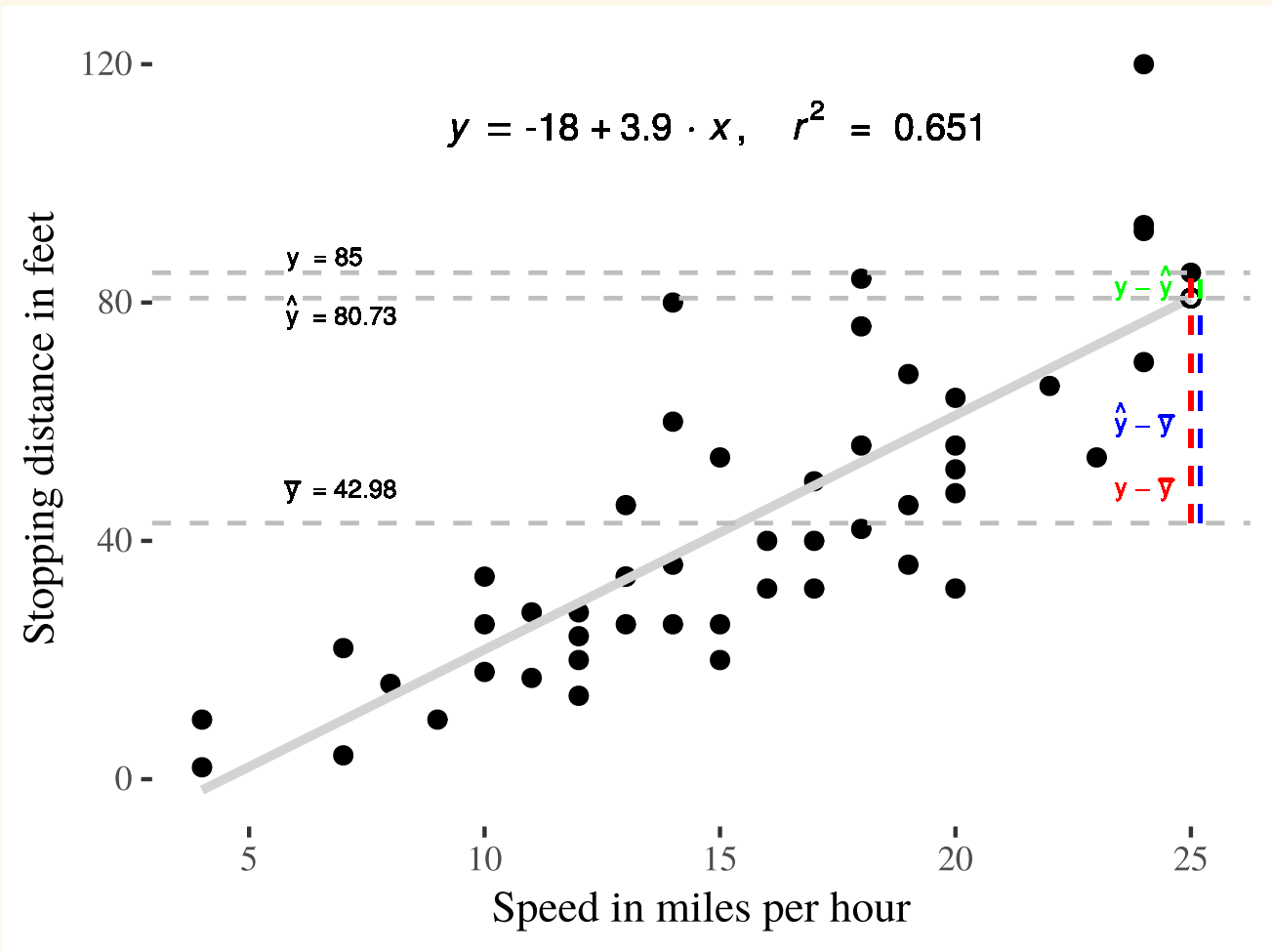
About 65% of the variation in stopping distances can be explained by the regression of stopping distances on speed.

Also ok:

About 65% of the variation in observed stopping distances can be explained by speed.

Example: ANOVA breakdown for (x,y) = (25,85)

```
ggplot(reg_points_cars, aes(x = speed, y = dist)) +  
  geom_point() +  
  theme(legend.position = "none") +  
  labs(x = "Speed in miles per hour",  
       y = "Stopping distance in feet") +  
  geom_smooth(method = "lm", se = FALSE, color = "l")  
  geom_point(aes(y = dist_hat[50], x = speed[50]),  
  geom_hline(yintercept=mean(cars$dist), col = 'green')  
  geom_text(aes(7,48, label=(paste(expression(hat(y  
  geom_hline(yintercept=reg_points_cars$dist_hat[50], col = 'grey')  
  geom_text(aes(6.64, 87, label=(paste(expression(y  
  geom_segment(aes(x = 25, xend = 25, y = mean(cars$  
  geom_text(aes(24,48, label=(paste(expression(y -  
  geom_segment(aes(x = 25.2, xend = 25.2, y = mean(c  
  geom_text(aes(24,60, label=(paste(expression(hat(  
  geom_segment(aes(x = 25.2, xend = 25.2, y = dist_h  
  geom_text(aes(24,83, label=(paste(expression(y -  
  geom_text(x = 15, y = 110, label = lm_eqn(cars),
```



Analysis of Variance (ANOVA) for SLR

$$SST = SSreg + SSR$$

- **SST: Total variation** Total sum of squares

$$SST = SSTot = \sum_{i=1}^n (y_i - \bar{y})^2 = (n - 1)s_y^2$$

- **SSR: Unexplained variation** Residual sum of squares

$$SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = (n - 2)\hat{\sigma}^2$$

- **SSreg: Explained variation** Regression sum of squares

$$SSreg = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

R-squared

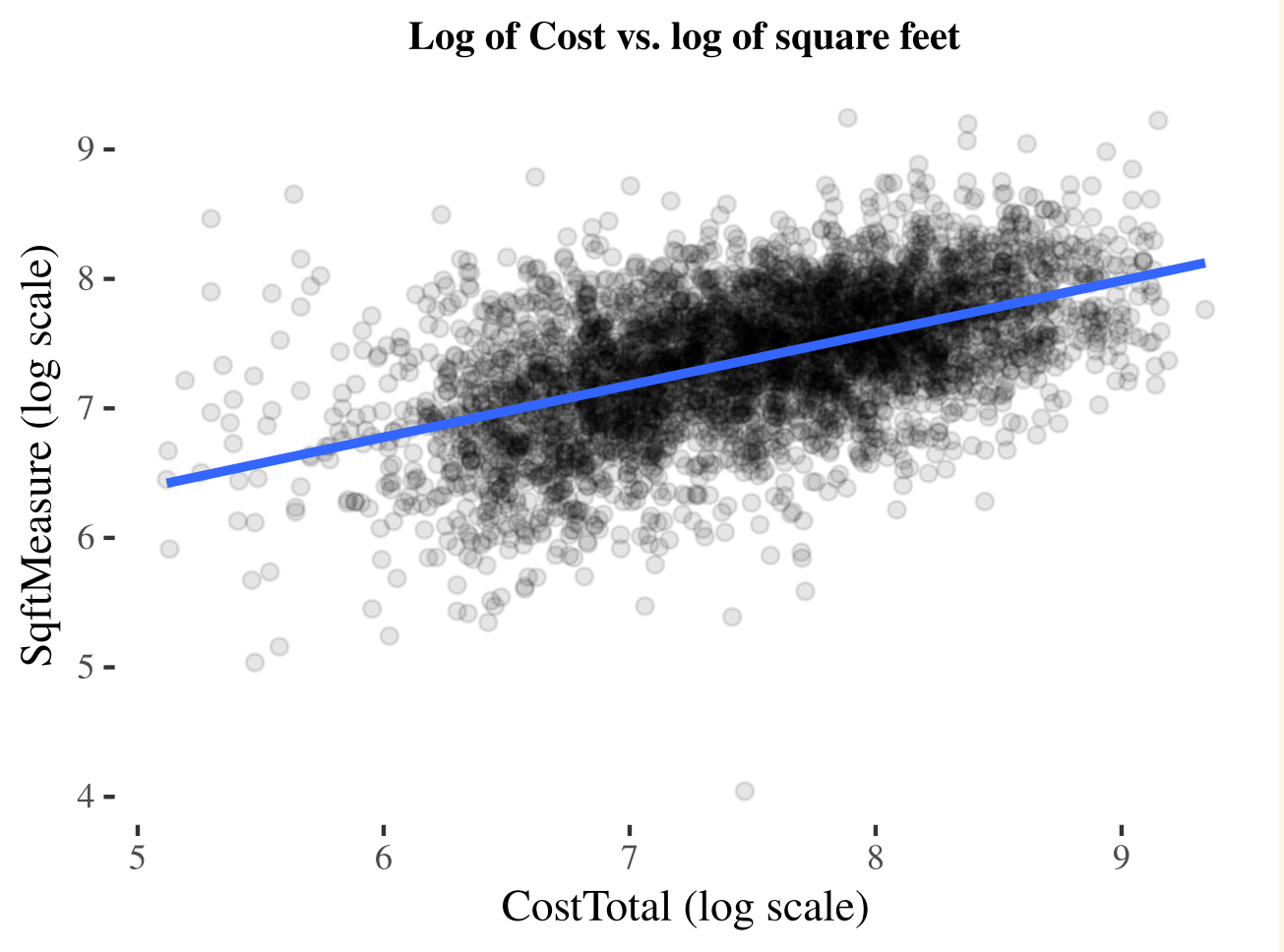
R-squared (R^2 or coefficient of determination) measures the proportion of variability observed in the response Y which can be explained by the regression of Y on x.

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{SS_{reg}}{SST}$$

- In terms of *unexplained* (residual) variation:

$$R^2 = 1 - \frac{\text{unexplained variation}}{\text{total variation}} = 1 - \frac{SSR}{SST} = 1 - \frac{(n-2)\hat{\sigma}^2}{(n-1)s_y^2}$$

```
energy <- energy %>%
  mutate(logCost = log(CostTotal),
         logSqft = log(SqftMeasure))
ggplot(energy, aes(x = logSqft, y = logCost)) +
  geom_point(alpha = .1) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Log of Cost vs. log of square feet",
       x = "CostTotal (log scale)",
       y = "SqftMeasure (log scale)") +
  theme(plot.title = element_text(hjust=0.5, size=9,
```



RECS

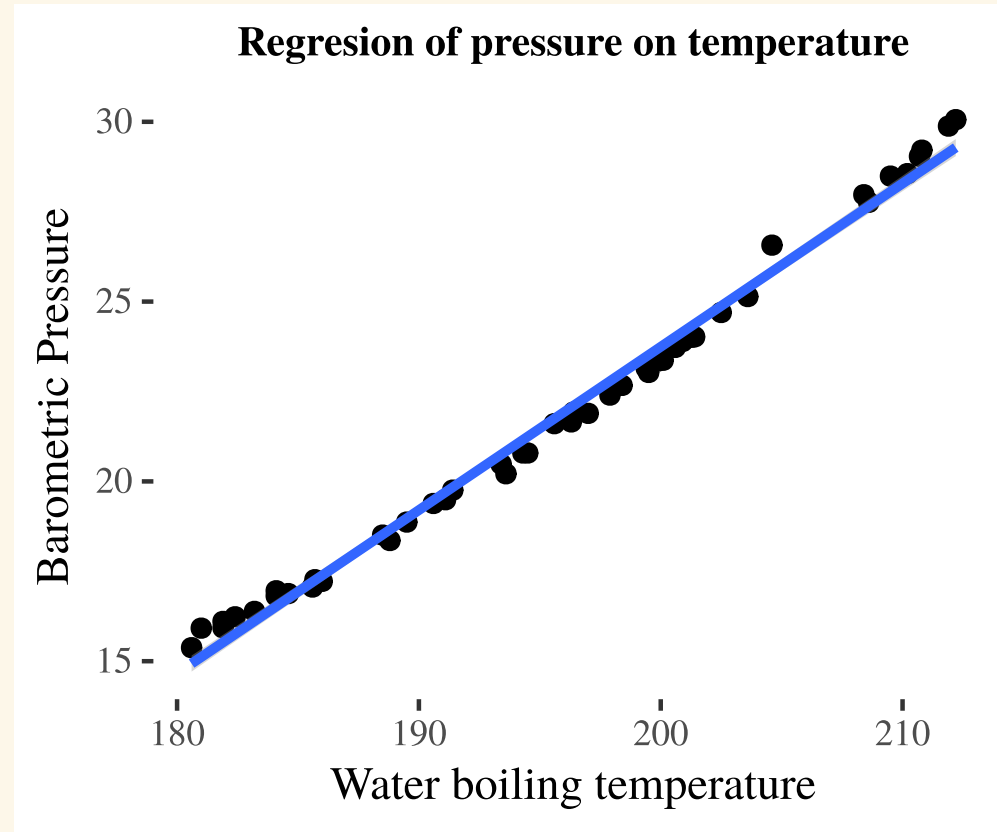
```
cost_lm <- lm(log(CostTotal) ~ log(SqftMeasure), data = energy)
summary(cost_lm)$r.squared
```

```
[1] 0.2787167
```

- The log of square footage only accounts for 27.9% of the variation in log of total energy cost.
- 72.1% of the variation in cost is unexplained by square footage
- A low R-squared like this does not mean that square foot is a worthless explanatory variable!
- We should explore a multiple regression model that includes more explanatory variables that could explain more of the variation in cost

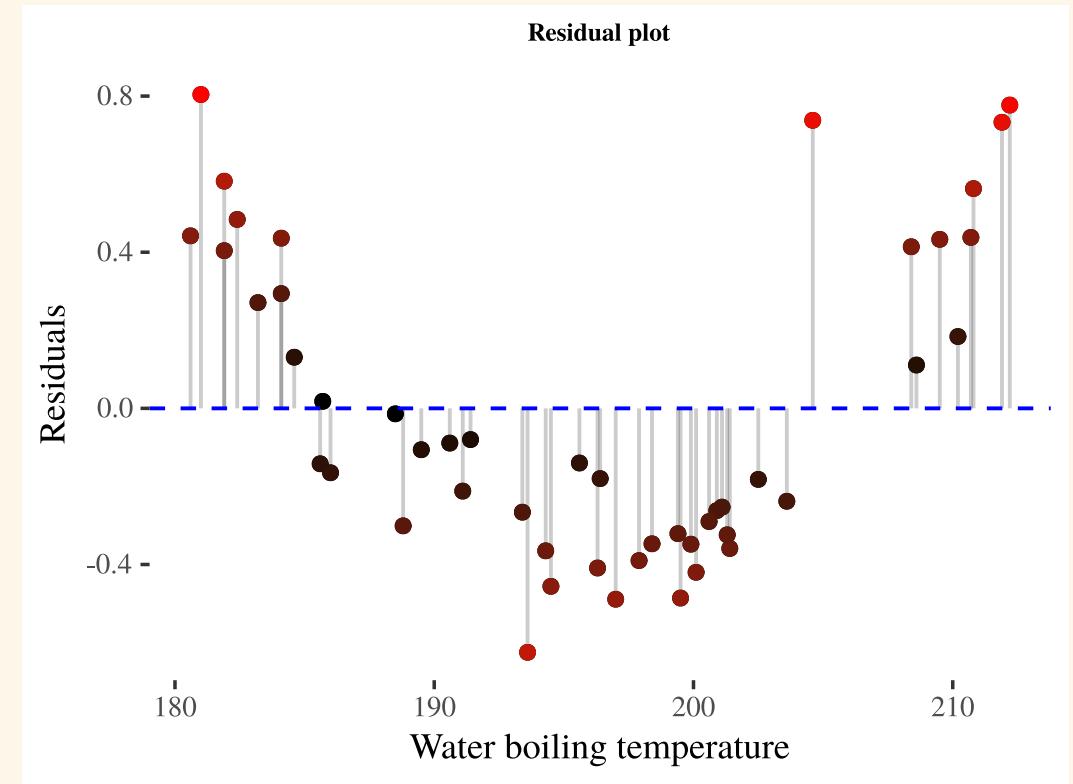
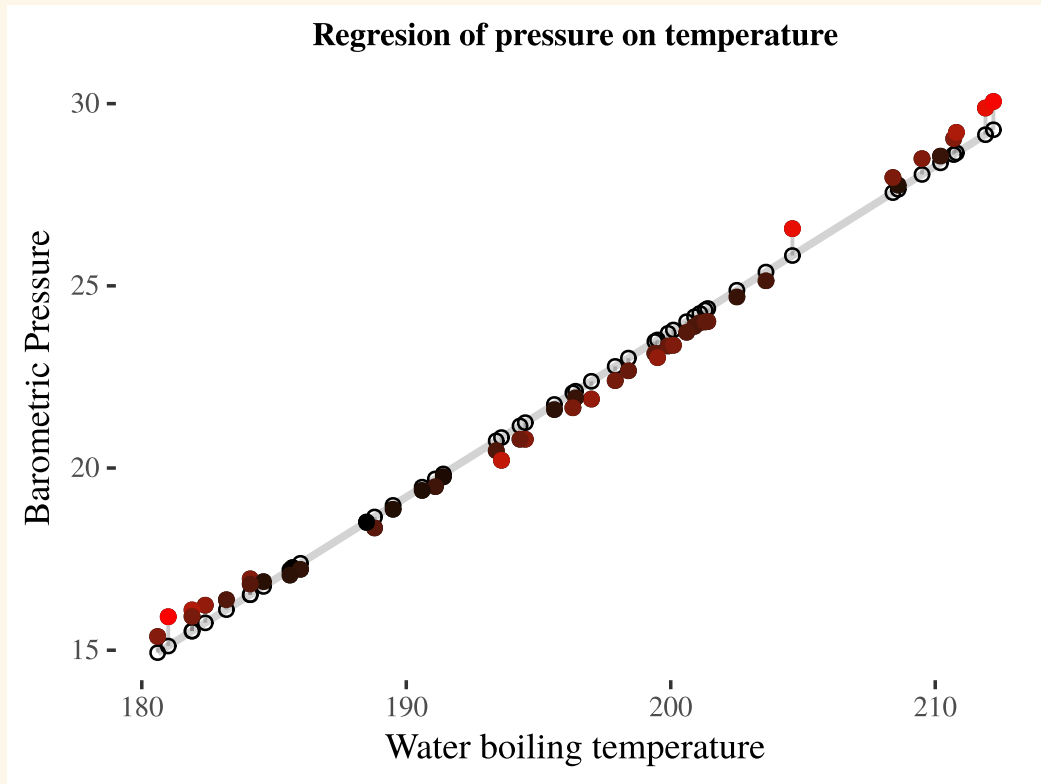
R-squared warning

Warning: Do not use only R-squared to assess the quality of a model unless you have verified that all model assumptions hold!!



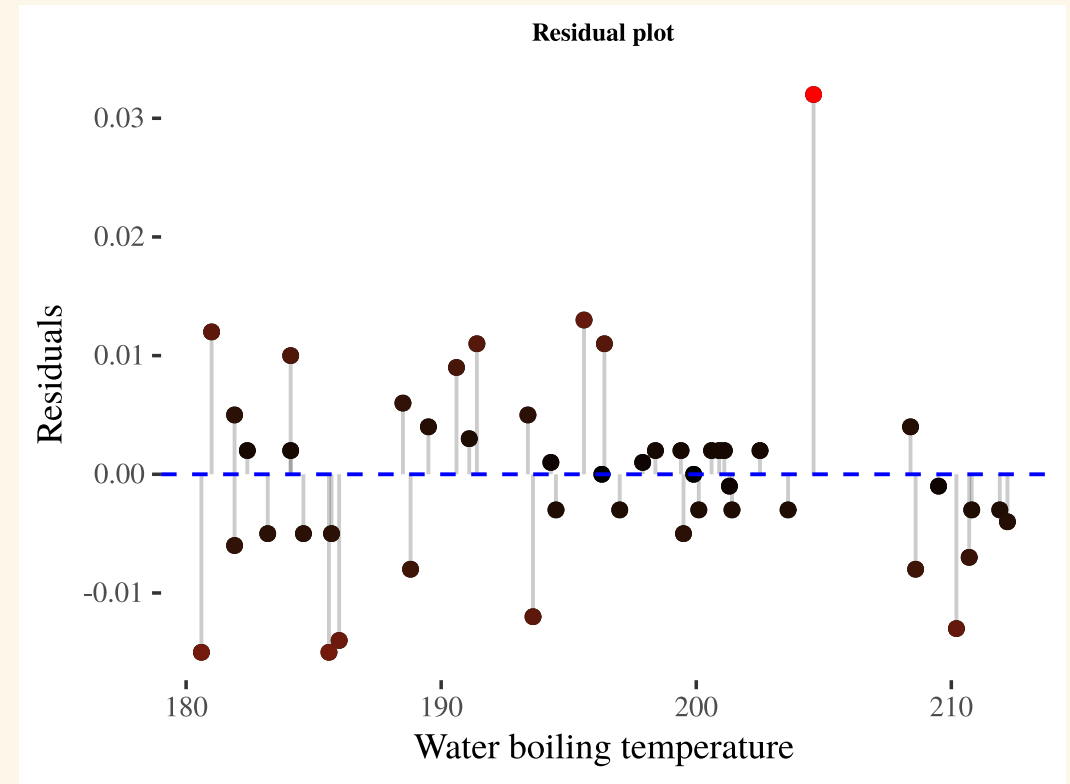
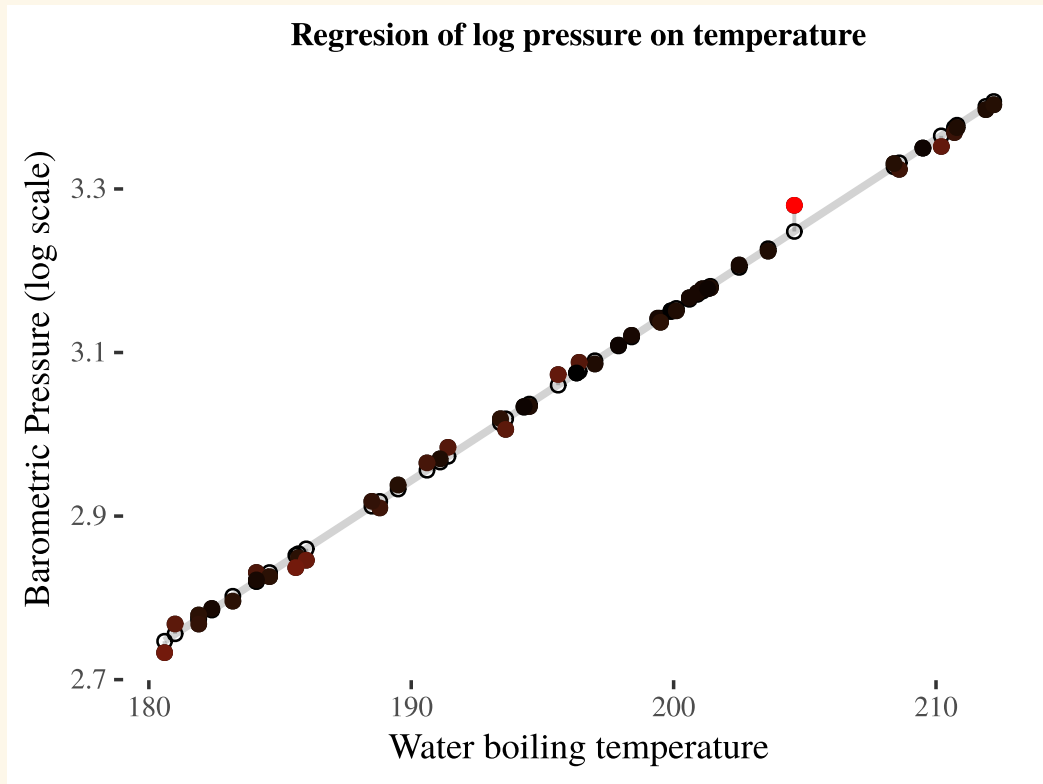
R-squared value is 99.8% but what about the assumptions?

R-squared warning



Warning: Do not use only R-squared to assess the quality of a model unless you have verified that all model assumptions hold!!

SLR of log pressure on temp does fit a SLR form



For this model, R-squared is 99.5 %. It doesn't matter that R-squared is lower, what matters is that the observed relationship now agrees with our SLR linearity assumption.

Your Turn 1

05:00



- Get the in class activity file from [moodle](#)
- We will further practice the concepts seen in the slides