PM 592
Regression Analysis for Public Health Data Science

Week 3

Regression I

Regression I

Regression Introduction

Variable Transformations – A Brief Overview

Model Assumptions

The Best Fit Line

Interpreting Regression Output

Hypothesis Testing of Model Parameters

Correlation

- ✓ Z-Scores
- ✓ Probability from a normal distribution
- ✓ Assessing Normality
- ✓ The Central Limit Theorem
- ✓ Other Distributions
- ✓ Z- and t-tests

Assume we have a continuous random variable *Y* and want to examine its association with another random variable *X*.

For each person we observe *X* and *Y*. Each person's observed values are denoted (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n), for person 1 through n.

Our goal is to build a model for Y as a function of X.

X and *Y* could be something like:

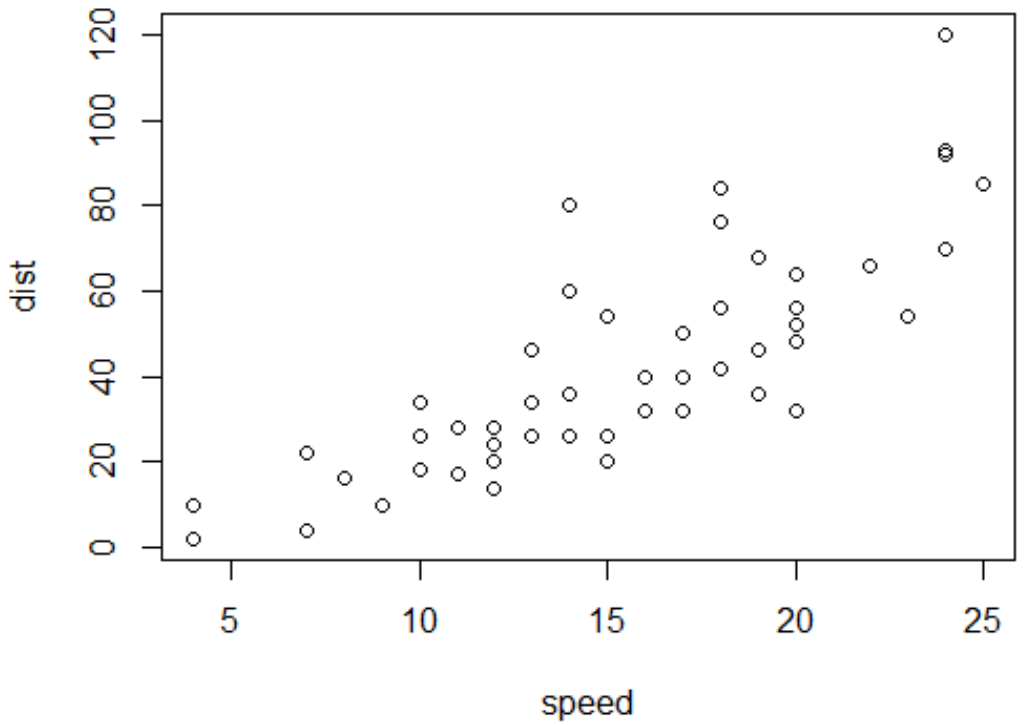
- X = height, Y = weight
- X = distance of a shop to a college, Y = revenue from selling vape cartridges
- X = # of drinks a customer has consumed, Y = amount gambled in a casino

Our first step is to **determine the model form**.

- Linear regression implies just that – the relationship between X and Y must be linear, or follow a straight line.
- X and Y could be related in other ways, but linear regression only detects the extent of the linear relationship.

Example

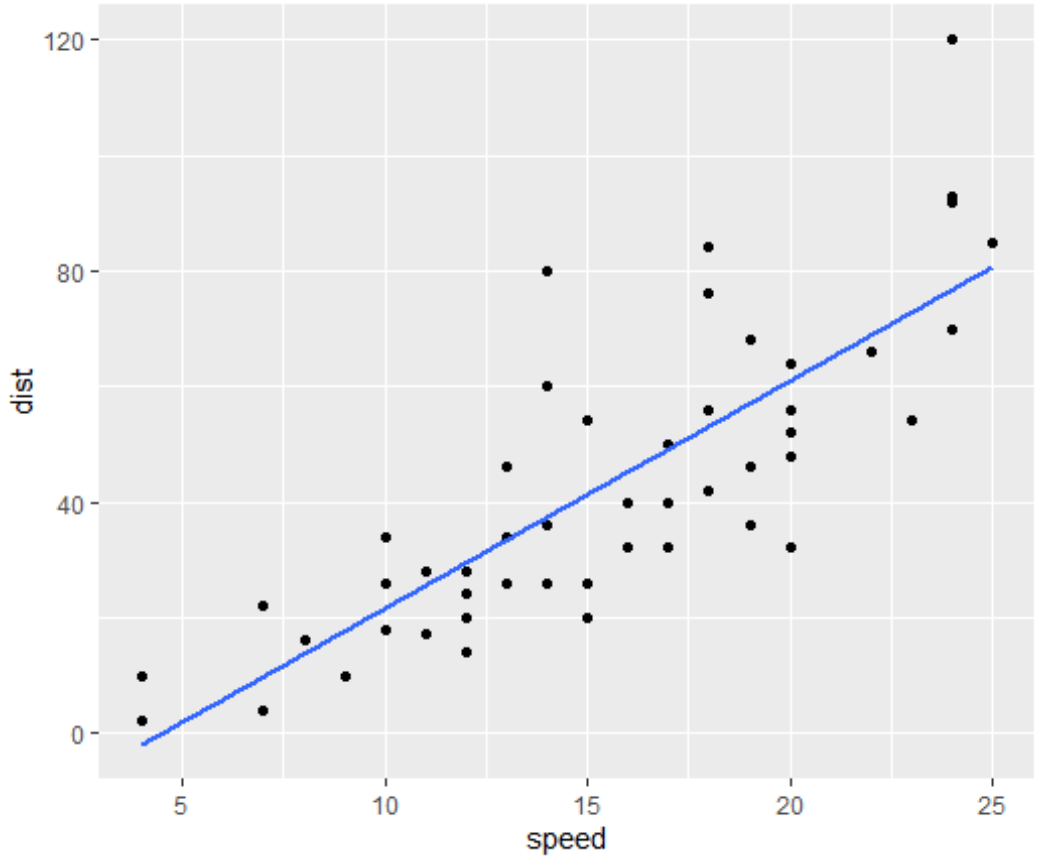
The relationship between car stopping distance (feet) and speed (mph)



Does this relationship appear linear?

It does appear that a line is “sufficient” to describe this relationship.

While we will go into more detail about how to examine linearity, this just *looks* linear to us.



The form of a regression line is:

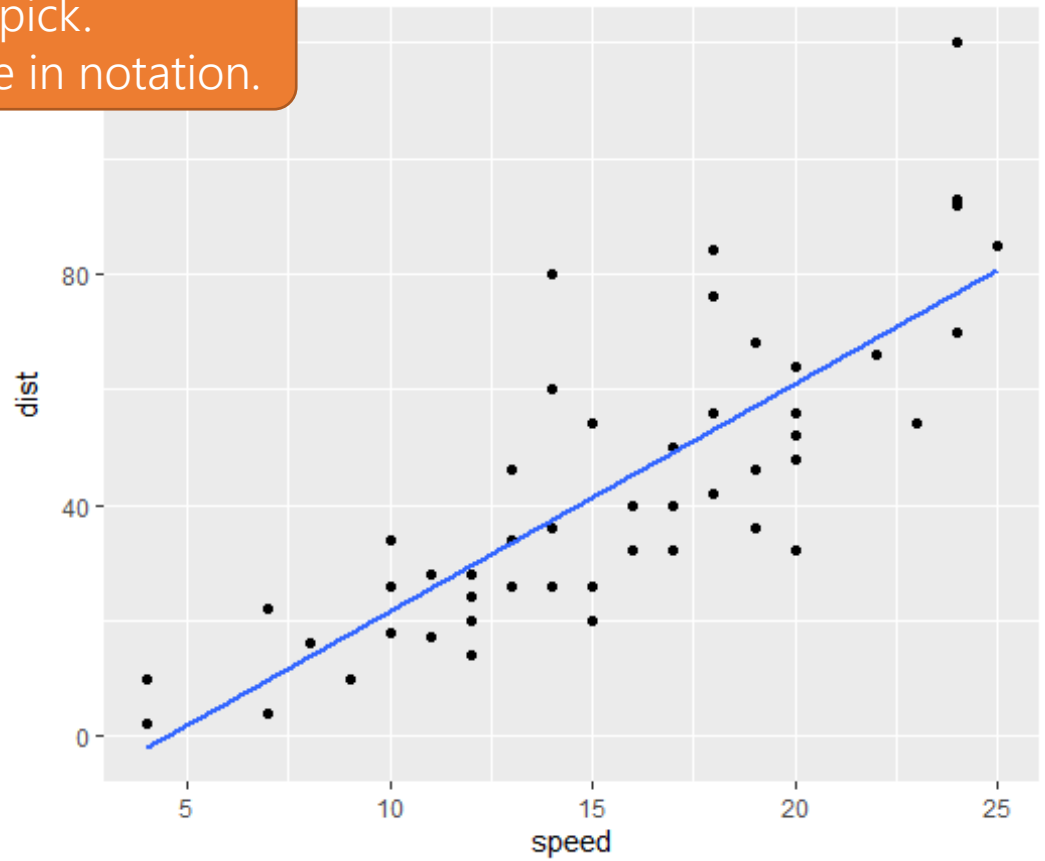
$$Y = \alpha + \beta X$$

$$Y = \beta_0 + \beta_1 X$$

Take your pick.
It's just a difference in notation.

The α value is the y-intercept (i.e., the value of y when x=0).

The β value is the slope (i.e., how much y increases when x increases by 1 unit).

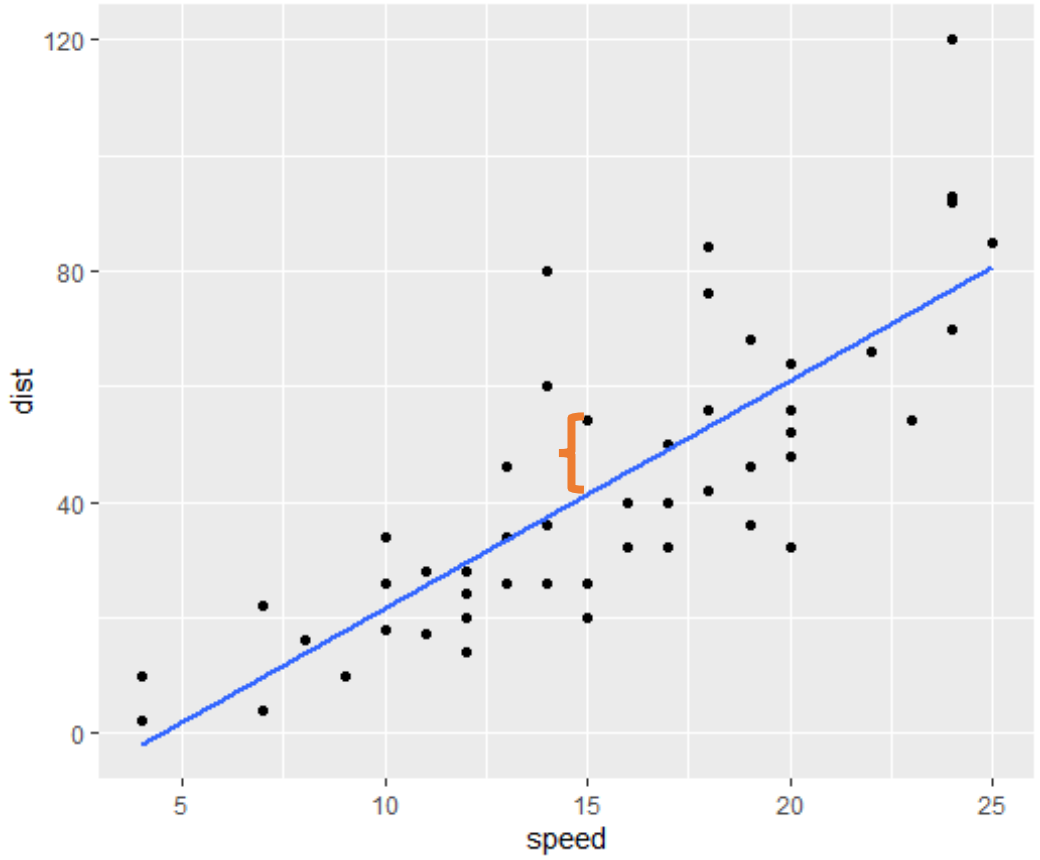


Prediction vs. Reality

Each observed measurement y_i does not fall on the regression line. There is always some error associated with these values; our model will never be perfect.

Our equation for somebody's actual Y value is therefore:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

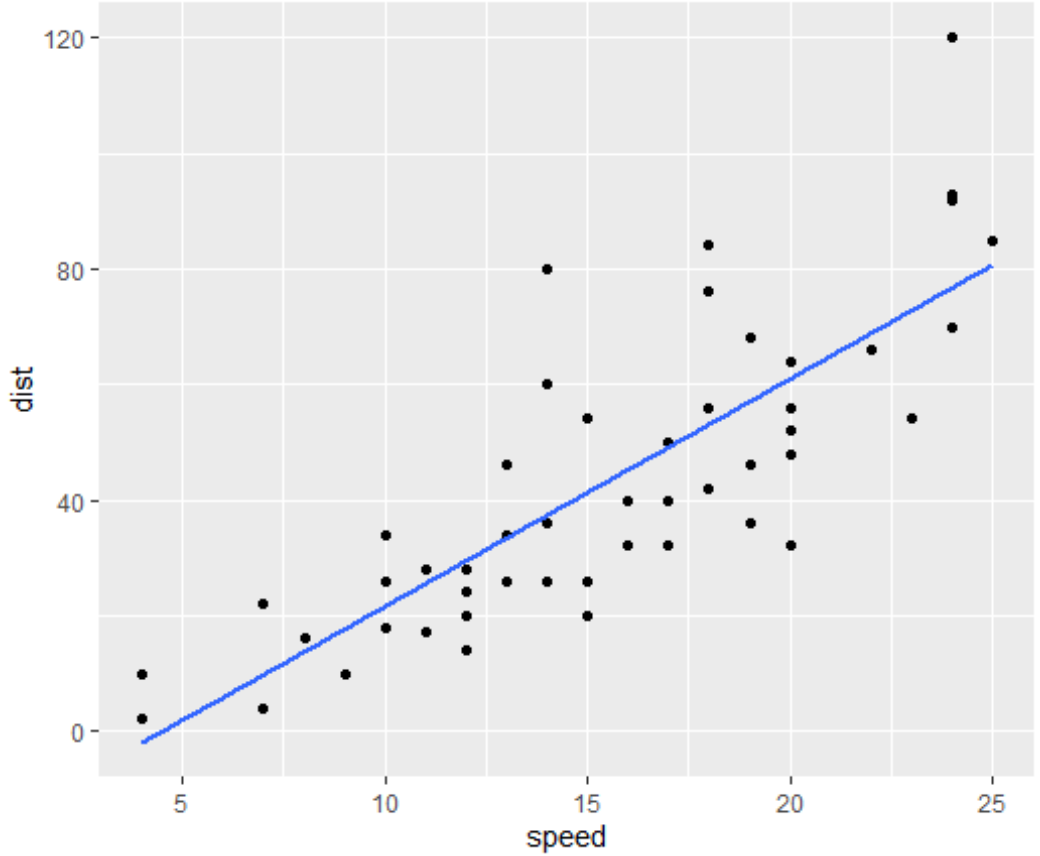


This is called the "residual." It's the part of a Y measurement that is not explained by the regression line.

Prediction vs. Reality

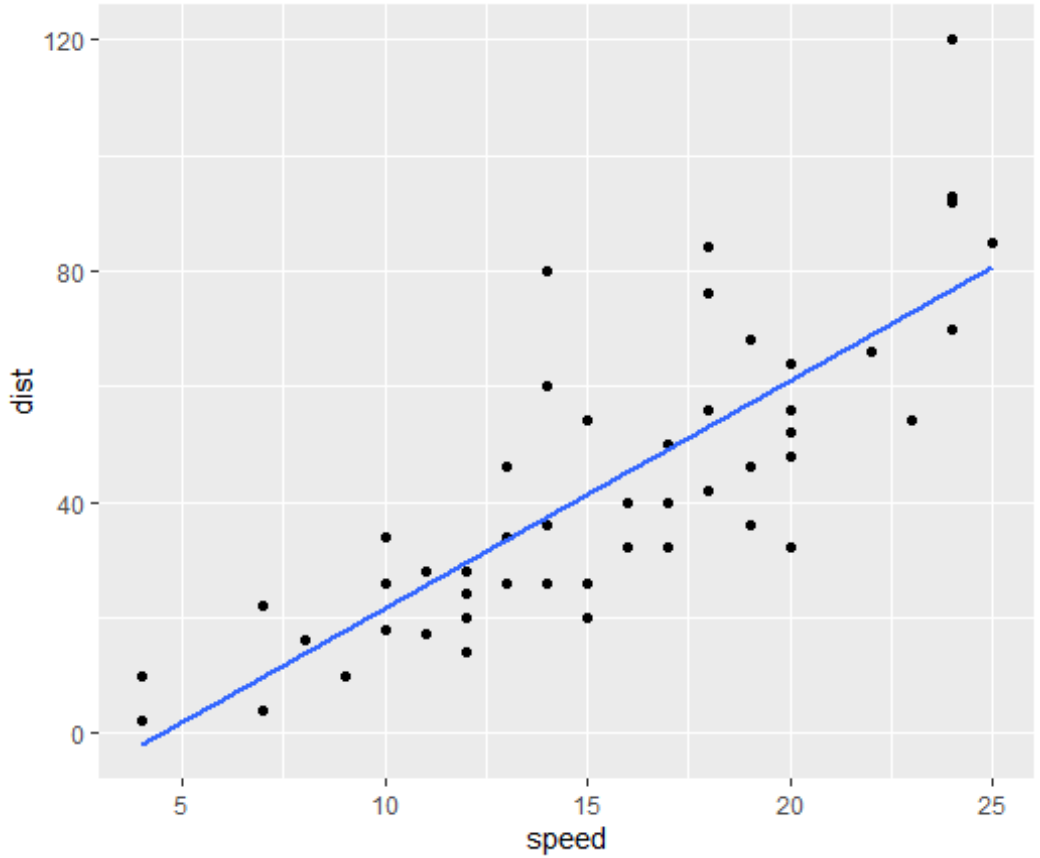
However we are more concerned with using this line to make predictions. Somebody's predicted value is given as:

$$\hat{Y} = \beta_0 + \beta_1 X$$



Let's run our first linear regression in R:

```
> lm(dist ~ speed, data = cars) %>%
+   summary()

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
```



Let's run our first linear regression in R:

```
> lm(dist ~ speed, data = cars) %>%
+   summary()

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
```

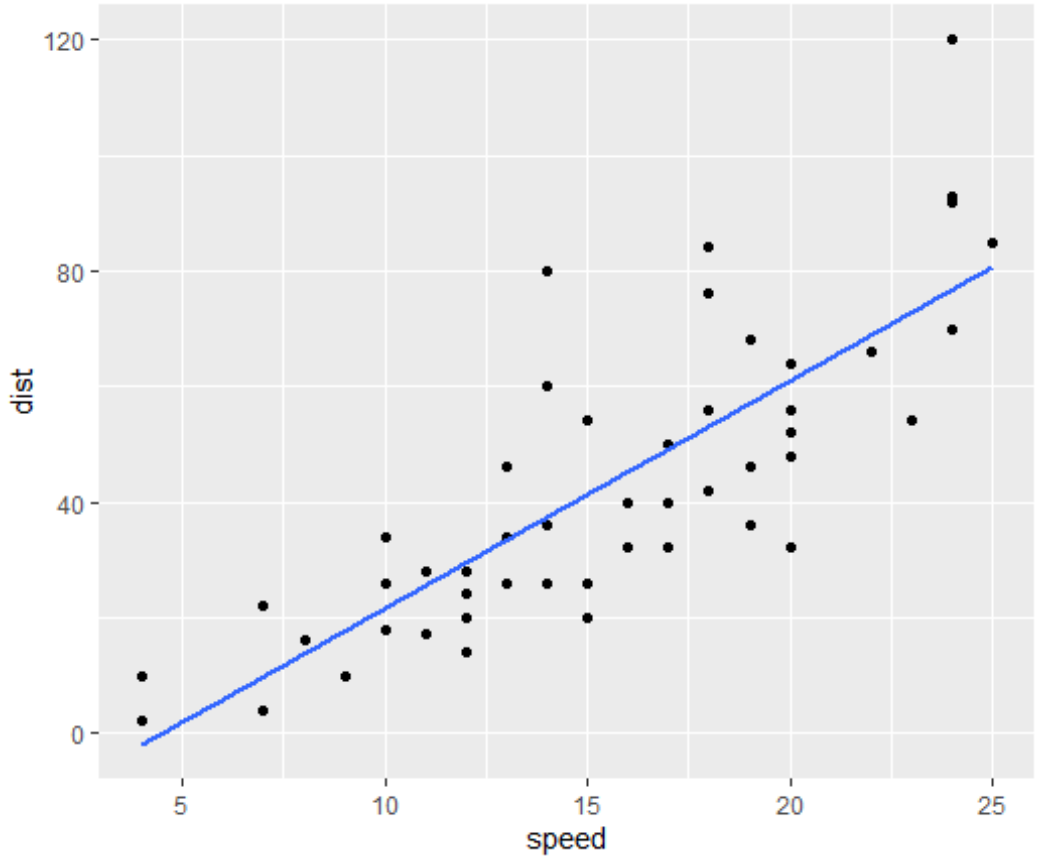
H_0: β_0=0

H_0: β_1=0

β_0

β_1, or β_{SPEED}

Regression coefficients follow a t-distribution on 1 df.



Q1a. What is the equation that describes the relationship between stopping distance and speed?

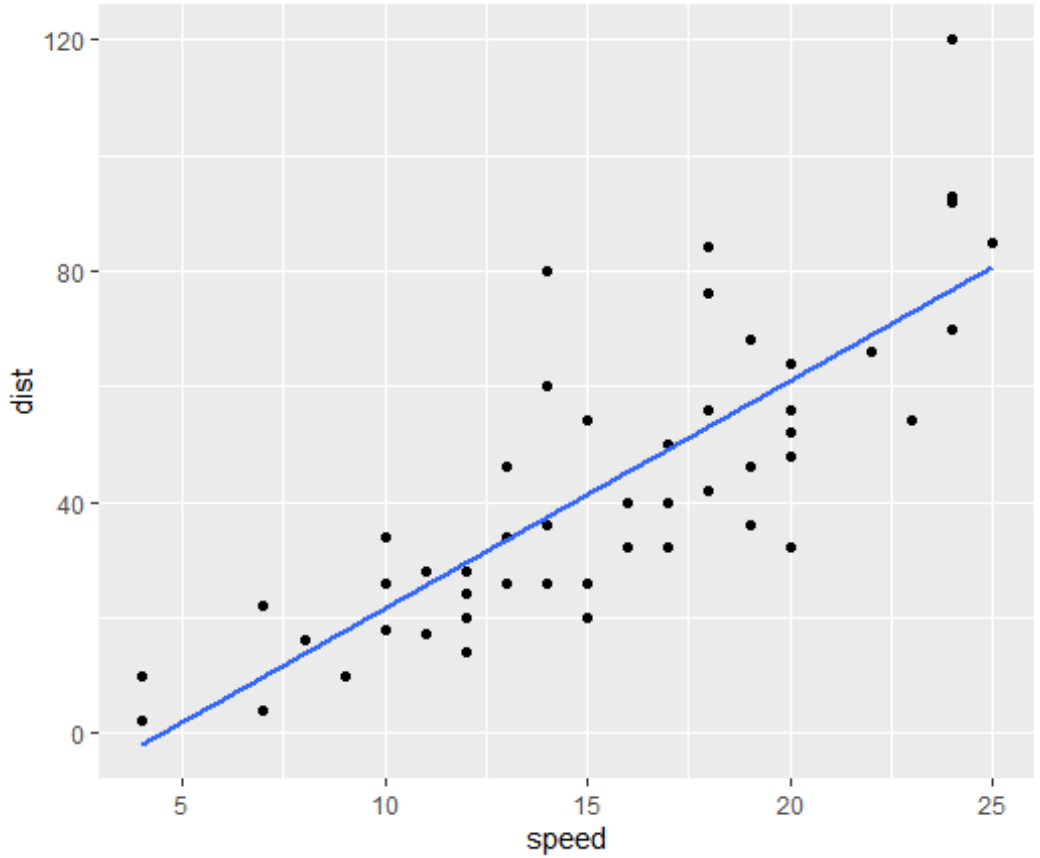
Q1b. Is stopping distance related to speed?

Q1c. What is the predicted stopping distance for a car that is not moving?

```
> lm(dist ~ speed, data = cars) %>%
+   summary()

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
```

Important Note

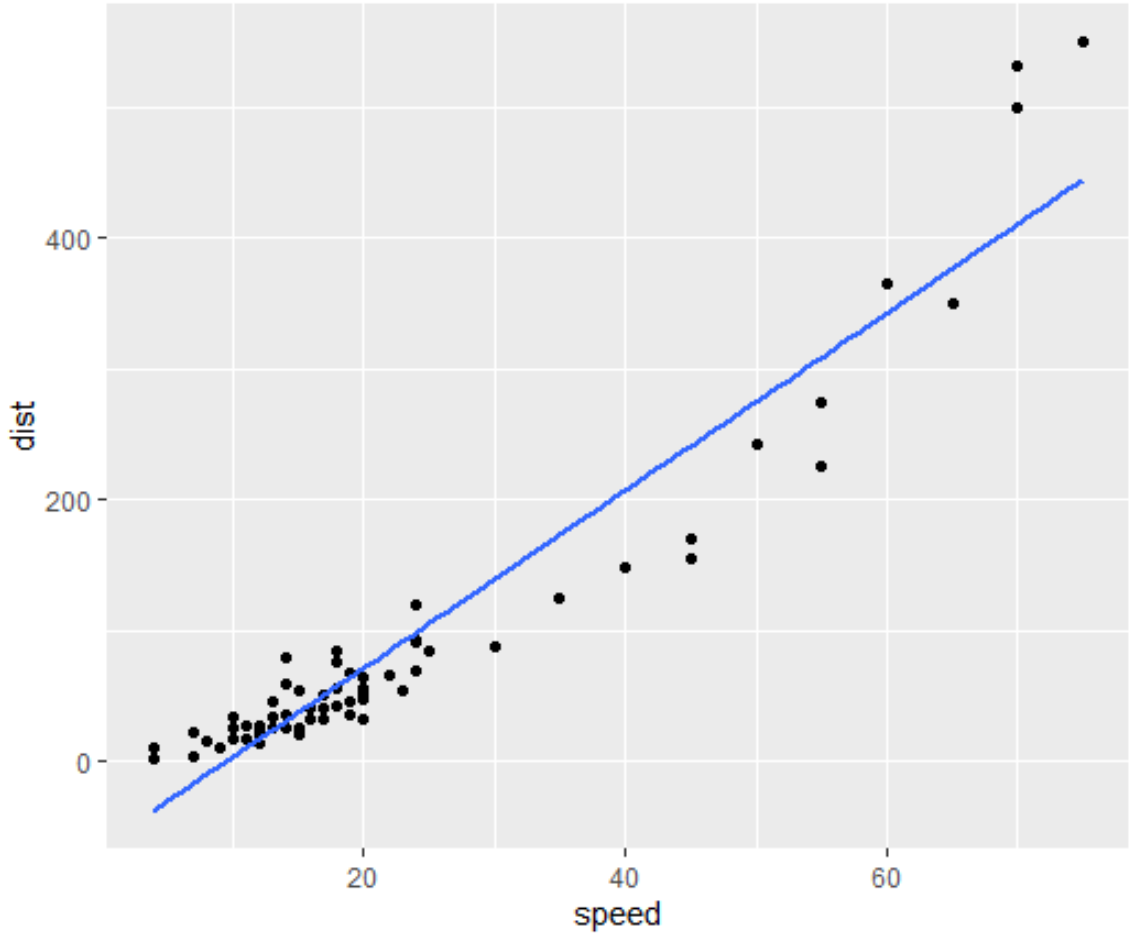
- When we perform a regression, we can only reasonably predict Y values across the range of X under study.
- Here, we can only draw an inference for speed values between 4 and 25mph.
- This is because the relationship may be nonlinear or may be different for X values not under study.



Suppose we study more values of speed – now we examine values from 4 to 90 mph.

Does the relationship appear linear after examining some more speed values?

- Well, Y does increase as X increases
- However adding a straight line to the plot shows us that a linear fit may not be completely appropriate.



Recap

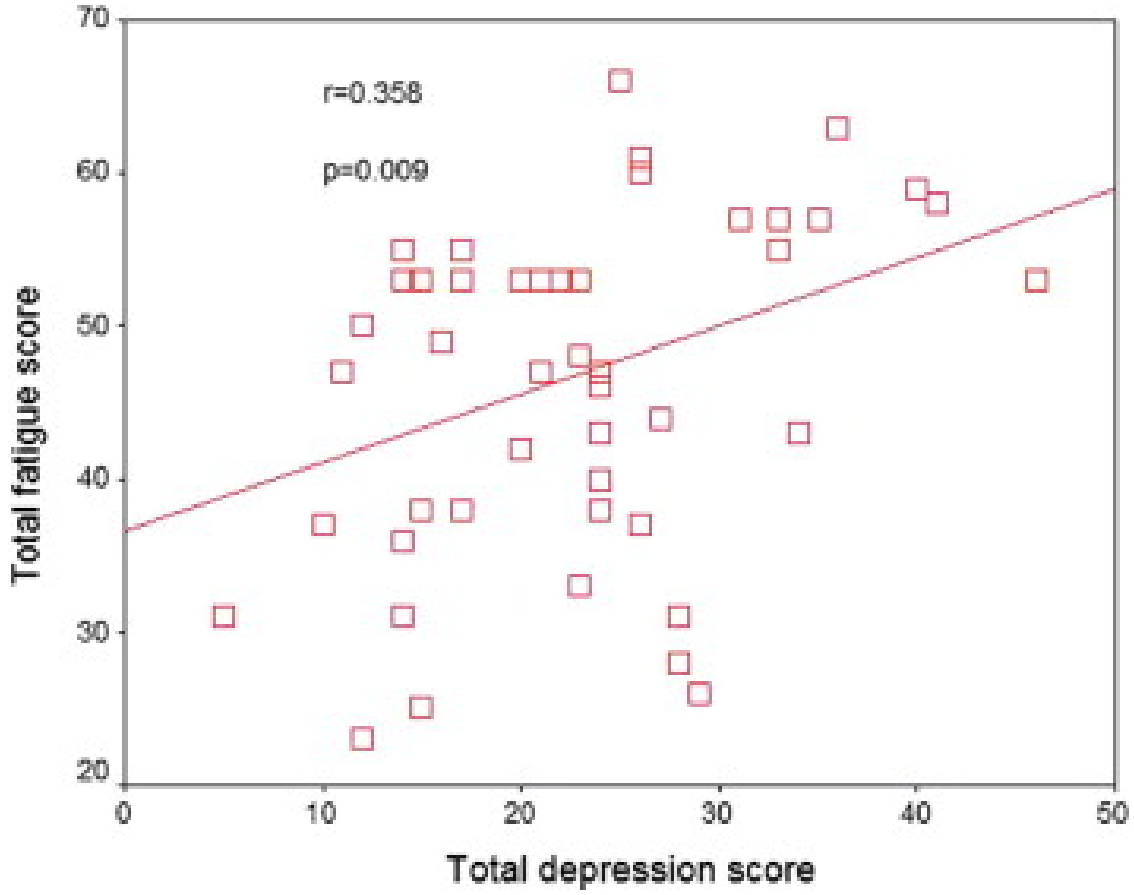
- The simplest form of regression models the linear relationship between X and Y and is called:
  - Ordinary least-squares (OLS) regression
  - Simple linear regression
  - General linear model
- Some relationships that appear linear may, in fact, not be linear upon examining a larger range of X values

Recap

- Write the form of the linear regression model
- Interpret the output of a linear model in R with respect to 1) the intercept and 2) the slope

Test Yourself

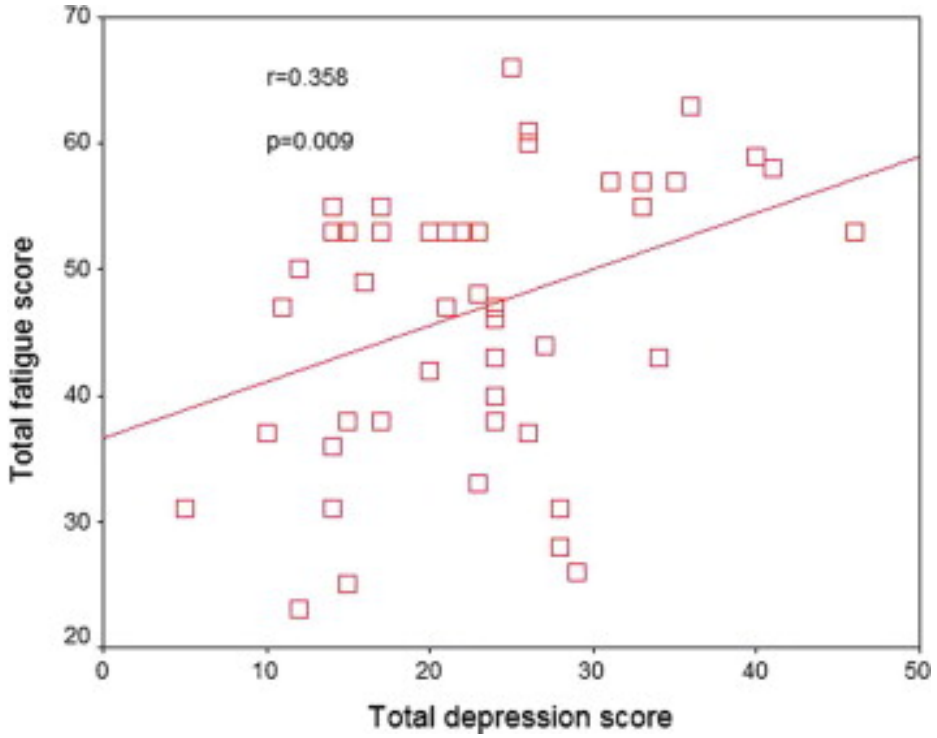
Estimate the value of the intercept from this regression relationship.



Test Yourself

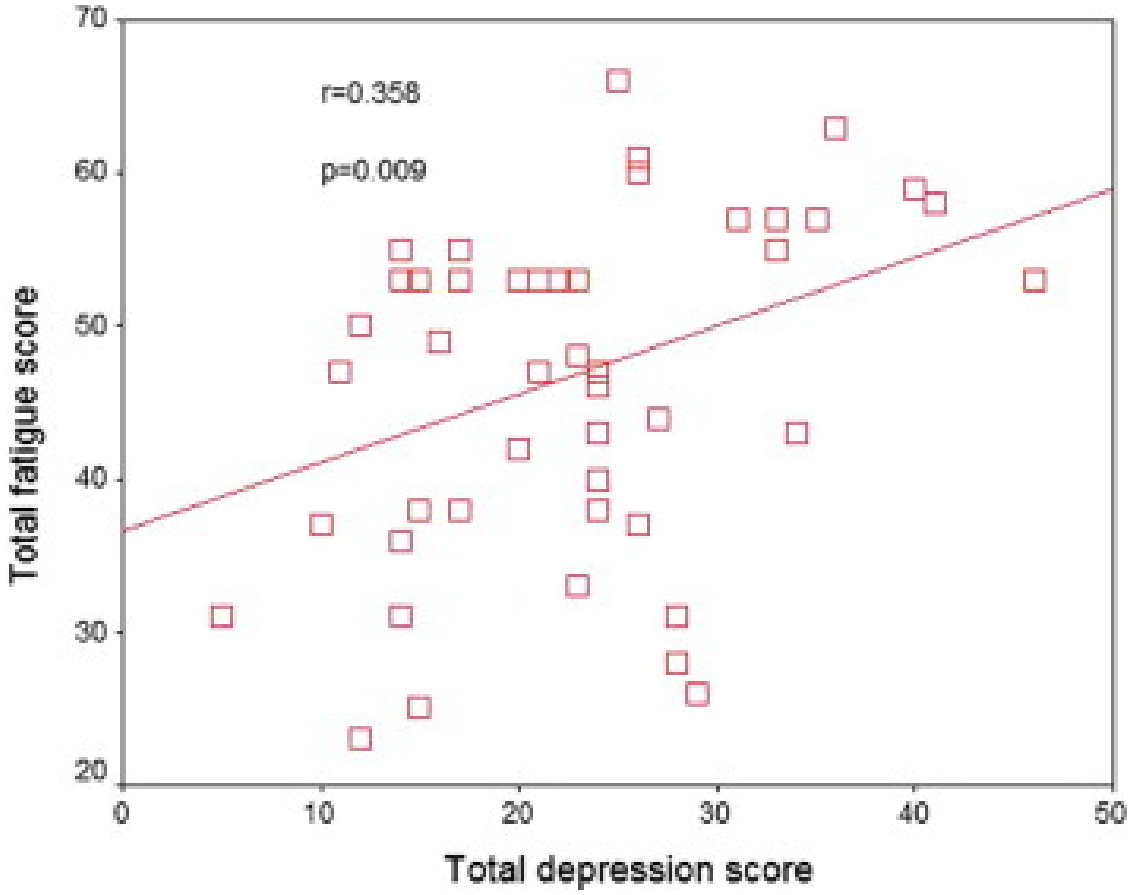
Estimate the value of the intercept from this regression relationship.

It looks like the regression line crosses the y-axis around 37 when X=0.



Test Yourself

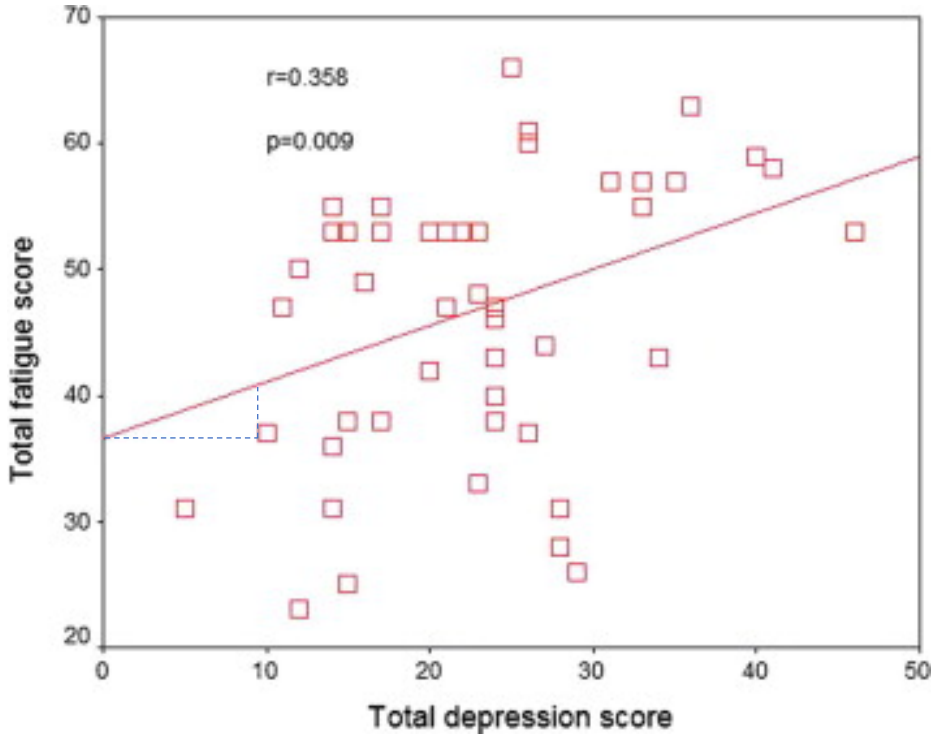
Estimate the value of the slope from this regression relationship.



Test Yourself

Estimate the value of the intercept from this regression relationship.

When X increases by 10, it looks like Y increases by ~4. So, we know that when X increases by 1, Y increases by ~0.4. My estimate of the slope would be 0.4.



When X and Y do not exhibit a linear relationship, there may be ways that we can make our variables conform to our analysis.

Option 1. Perform nonlinear regression.

- Useful we know the model form a priori (e.g., exponential, decay, predator-prey model, etc.)
- Model fitting can be difficult with this approach

Option 2. Perform linear regression and adjust the model/data to make it fit.

- Many techniques are available for linear regression
- Can be useful when the departure from linearity is not too extreme

Tukey's Ladder of Transformations

Find some transformation of the Y variable (Y*) that makes the relationship between X and Y* linear.

λ is the index of the transformation

λ	-2	-1	0	1/2	1	2	3
y^*	$1/y^2$	$1/y$	$\log(y)$	$y^{1/2}$	y	y^2	y^3

$\lambda < 1$: Reducing. Pulls in the right tail of a positively skewed distribution.

$\lambda = 1$: No transformation.

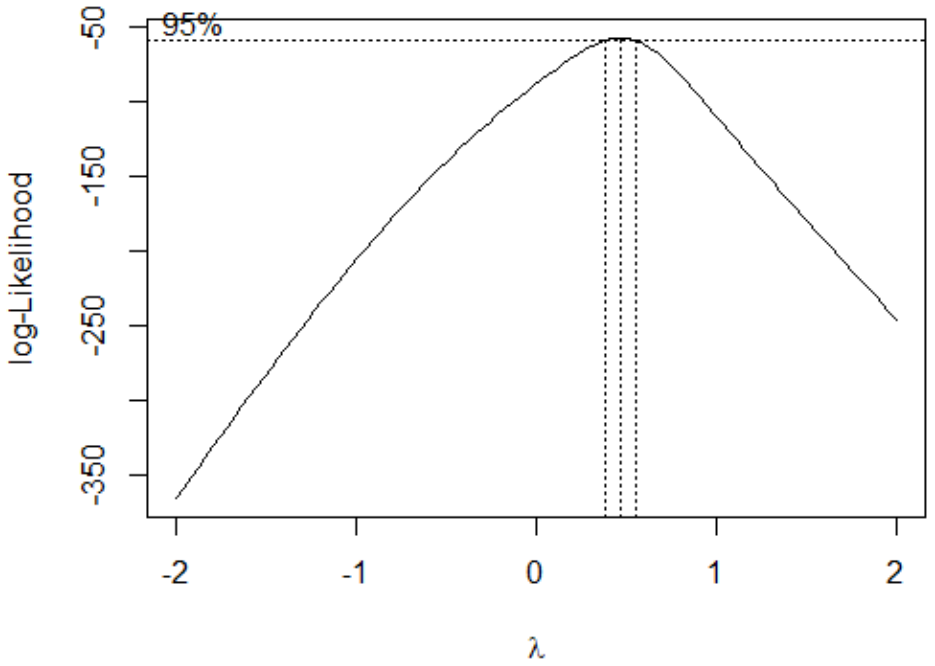
$\lambda > 1$: Expanding. Pulls in the left tail of a negatively skewed distribution.

Box-Cox Transformation

A procedure for finding the best transformation for the Y variable, given a linear regression.

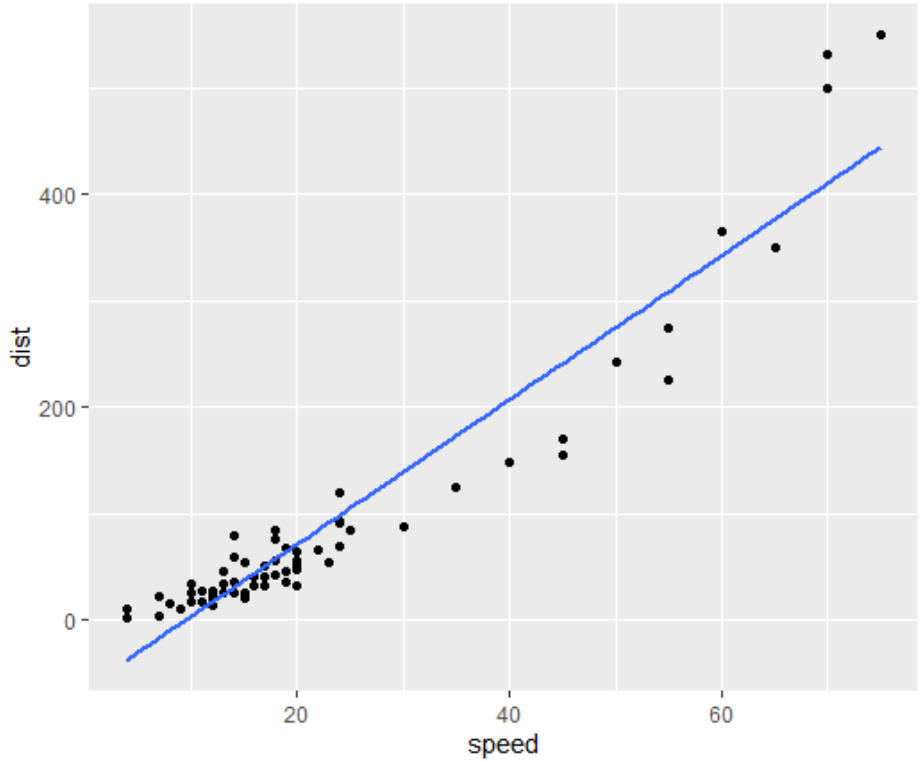
Here, the procedure finds the best lambda value for the relationship.

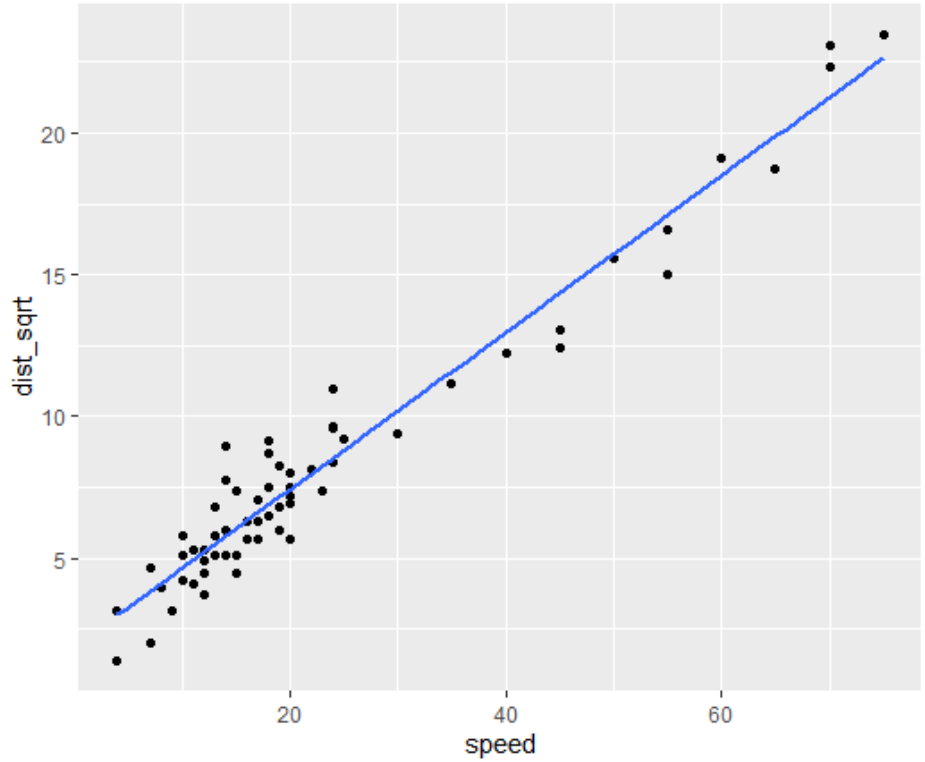
This is the value of lambda that maximizes the log-likelihood (more on that later in the course)



Q1. Which transformation is best for Y?

As we see below, the better linear relationship occurs between the square-root of distance and speed.

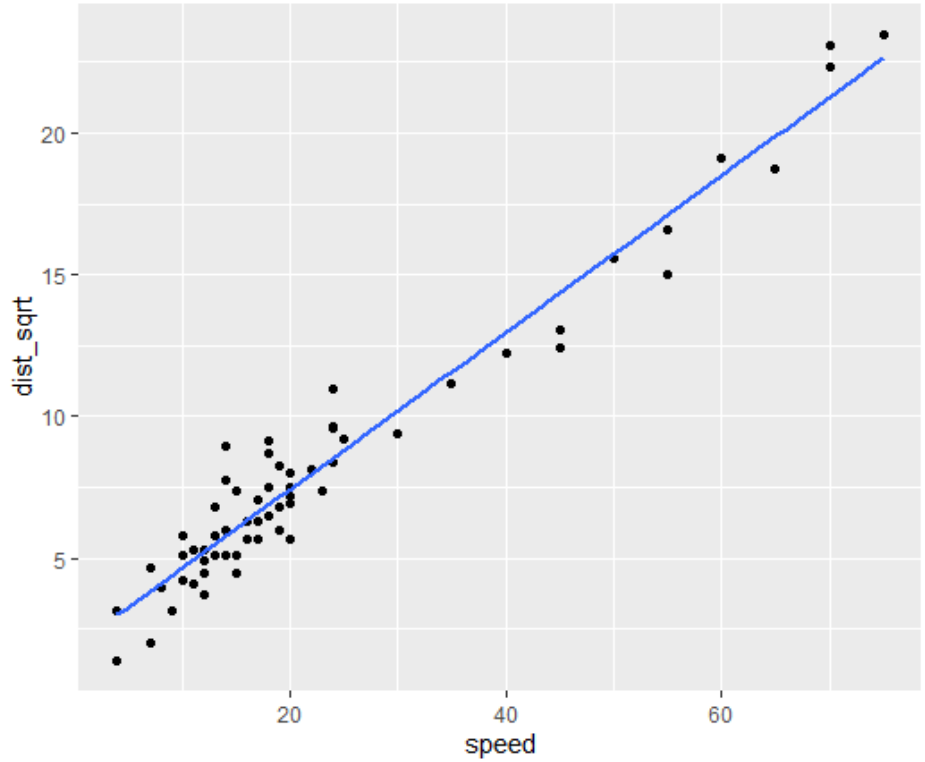




While a transformation of Y is straightforward, the interpretation of the model isn't quite as simple.

When speed increases by 1 mph, the *square root of* stopping distance increases by 0.28.

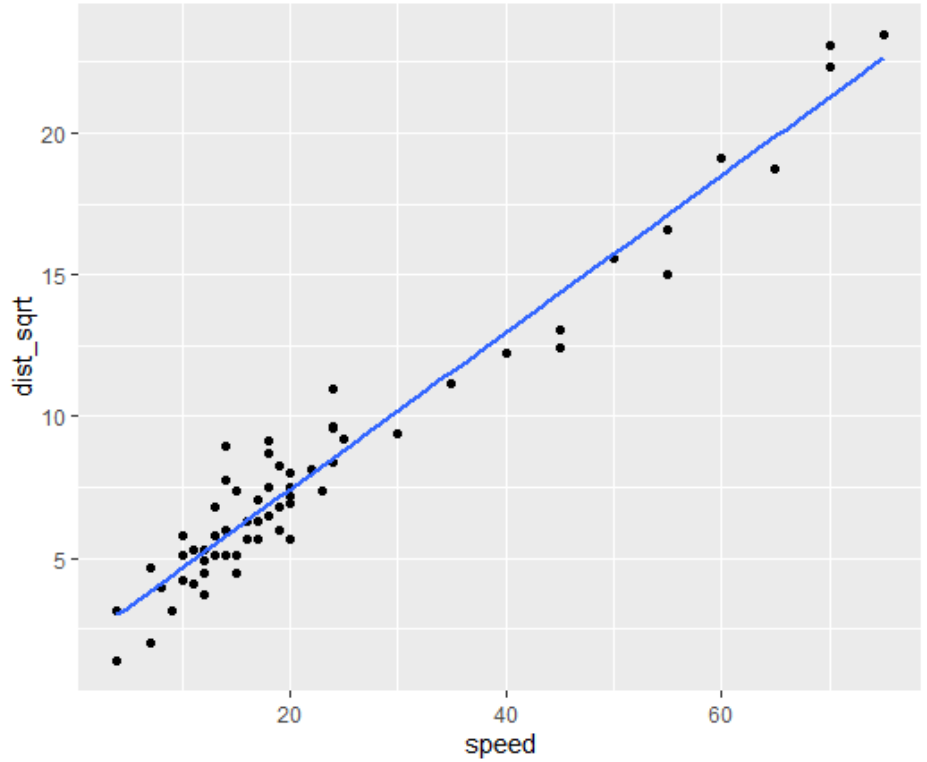
```
> lm(dist_sqrt ~ speed, data = carstot) %>%
+   summary()

Call:
lm(formula = dist_sqrt ~ speed, data = carstot)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1057 -0.7780 -0.1337  0.6287  3.1834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.887082   0.242072   7.796 9.89e-11 ***
speed       0.276702   0.008362  33.092  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.145 on 61 degrees of freedom
Multiple R-squared:  0.9472,     Adjusted R-squared:  0.9464
F-statistic:  1095 on 1 and 61 DF,  p-value: < 2.2e-16
```



When making predictions, we must take this transformation into account.

Q2. What is the predicted stopping distance for a car travelling 50 mph?

```
> lm(dist_sqrt ~ speed, data = carstot) %>%
+   summary()

Call:
lm(formula = dist_sqrt ~ speed, data = carstot)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1057 -0.7780 -0.1337  0.6287  3.1834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.887082   0.242072   7.796 9.89e-11 ***
speed       0.276702   0.008362  33.092  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.145 on 61 degrees of freedom
Multiple R-squared:  0.9472,    Adjusted R-squared:  0.9464
F-statistic:  1095 on 1 and 61 DF,  p-value: < 2.2e-16
```



Recap

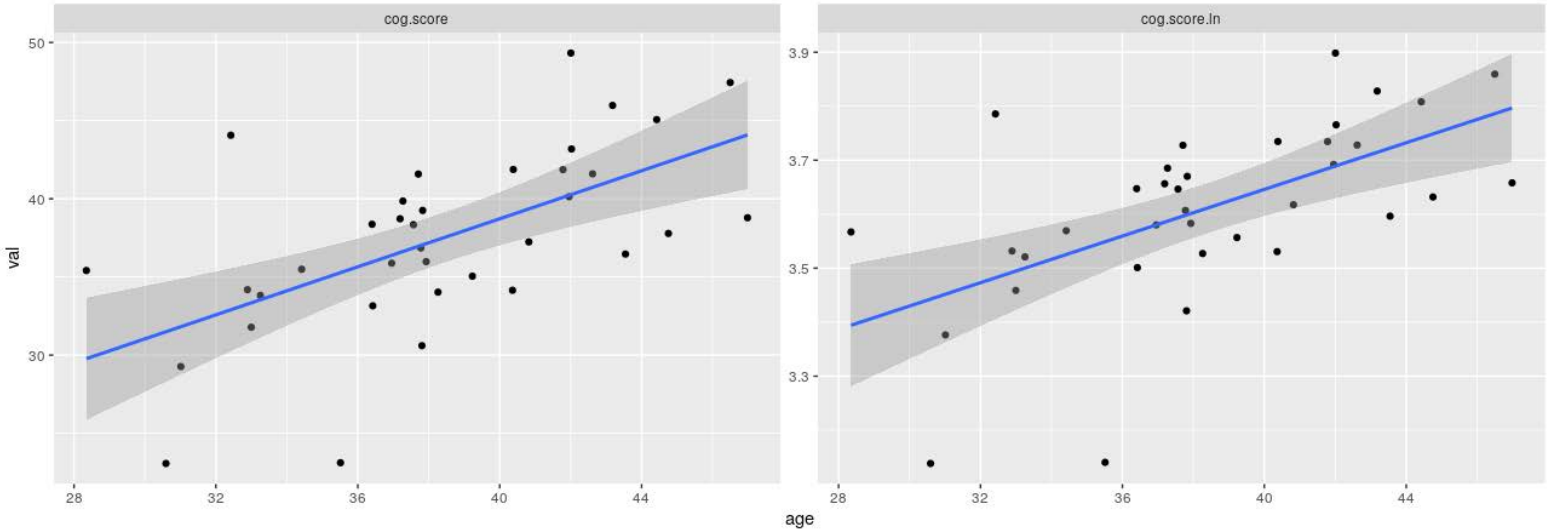
- When the relationship between X and Y appears nonlinear, the relationship can (sometimes) still be modeled in the linear regression framework by transforming the Y variable
- The Box-Cox procedure can determine the best transformation of Y

Recap

- Determine the best (if any) transformation of Y when the relationship between X and Y appears nonlinear
- Explain the pros and cons of transforming the Y variable

Test Yourself

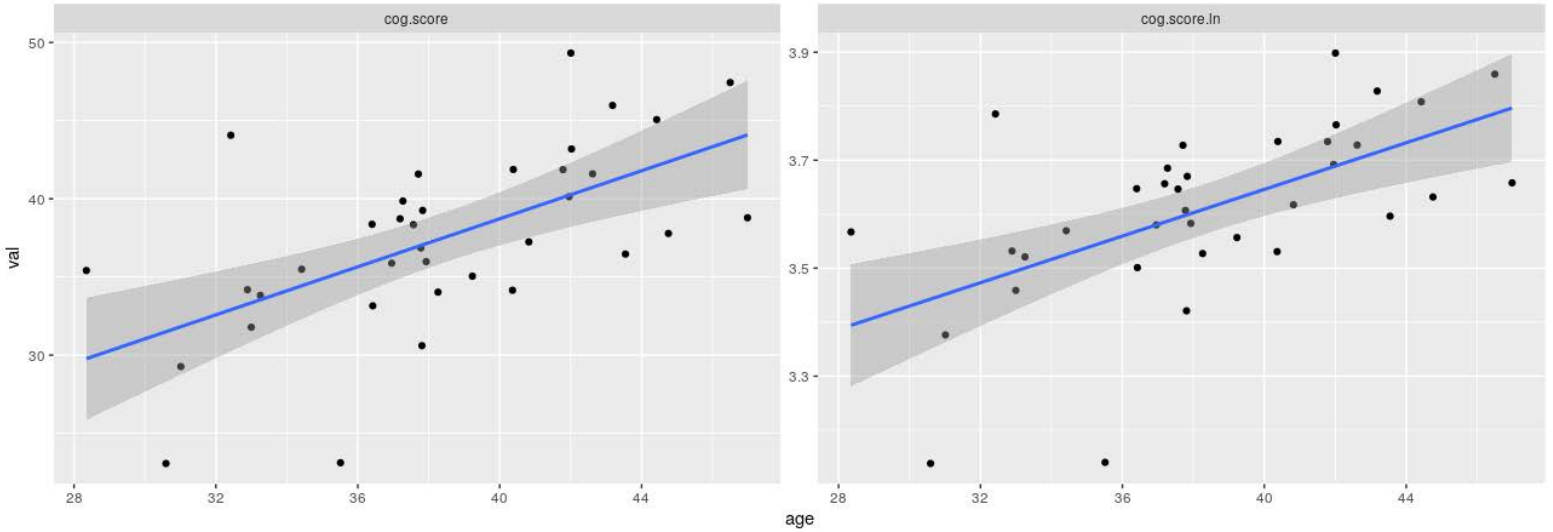
Consider the regression of cognitive score on age. Based on this output, would you use the regression with the raw cognitive score (left) or the log-transformed cognitive score (right)?



Test Yourself

Consider the regression of cognitive score on age. Based on this output, would you use the regression with the raw cognitive score (left) or the log-transformed cognitive score (right)?

It doesn't look like the log transformation really helps explain the relationship. I'd use the raw scores, for simplicity.



A linear regression model is only valid if it satisfies certain assumptions. While mild violations of the assumptions are common, more serious violations can be cause for concern.

There are four assumptions of **LINEar** regression:

1. **L**inearity
2. **I**ndependence
3. **N**ormality
4. **E**quality of variances (“homoscedasticity”)

Linearity

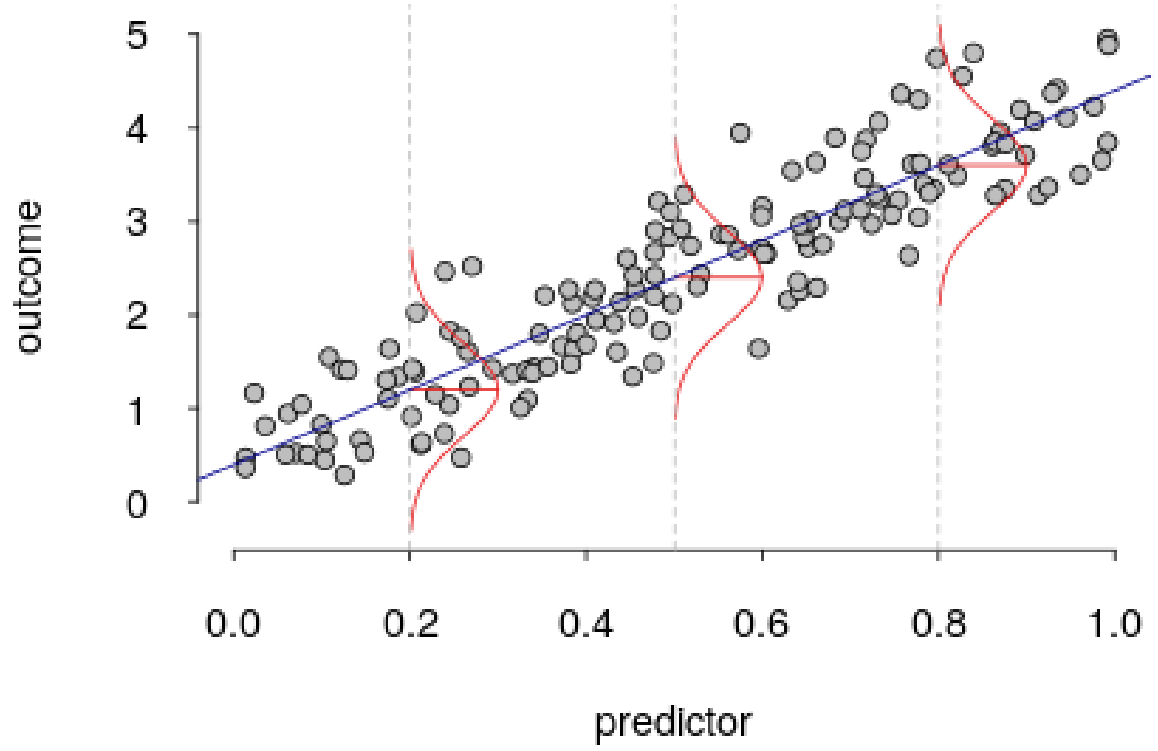
This one seems the most straightforward. However, let's think about it in greater detail…

If a relationship between X and Y is truly linear, then we can say the following:

$$\mu_{Y|X} = \beta_0 + \beta_1 x$$

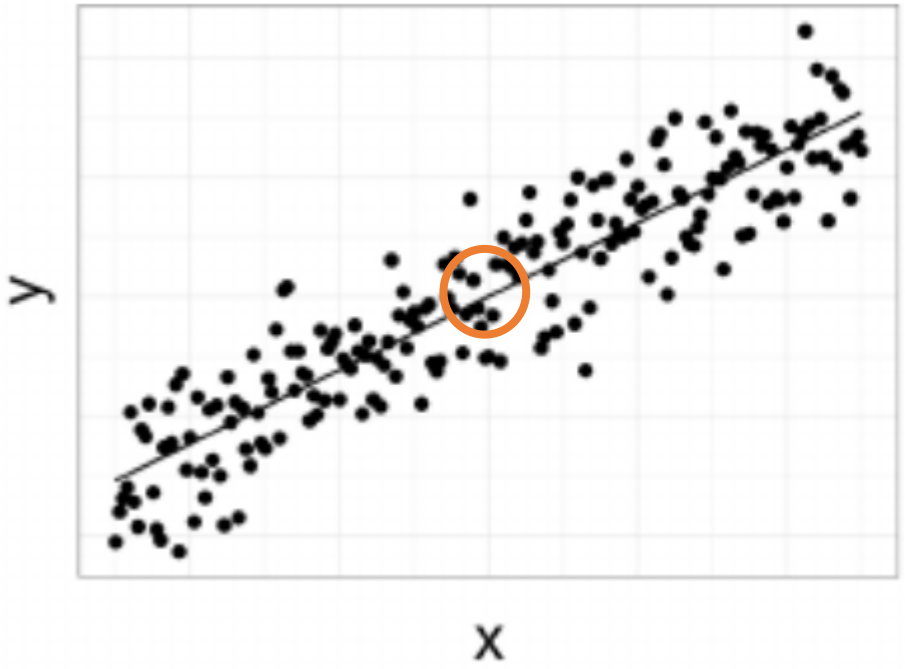
Linearity

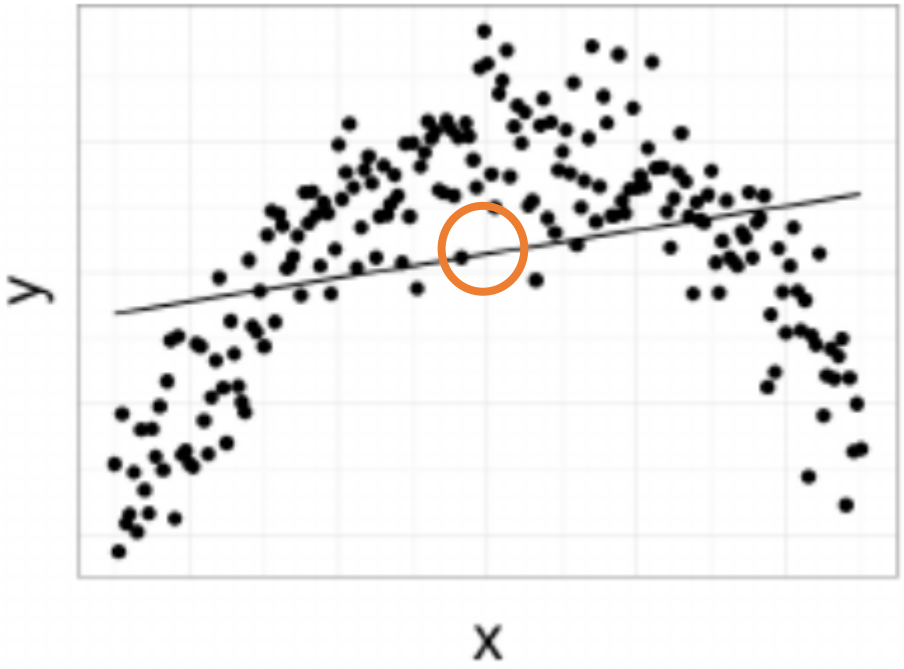
That is, our regression line goes through the mean of Y across all X values.



Take the following two examples.

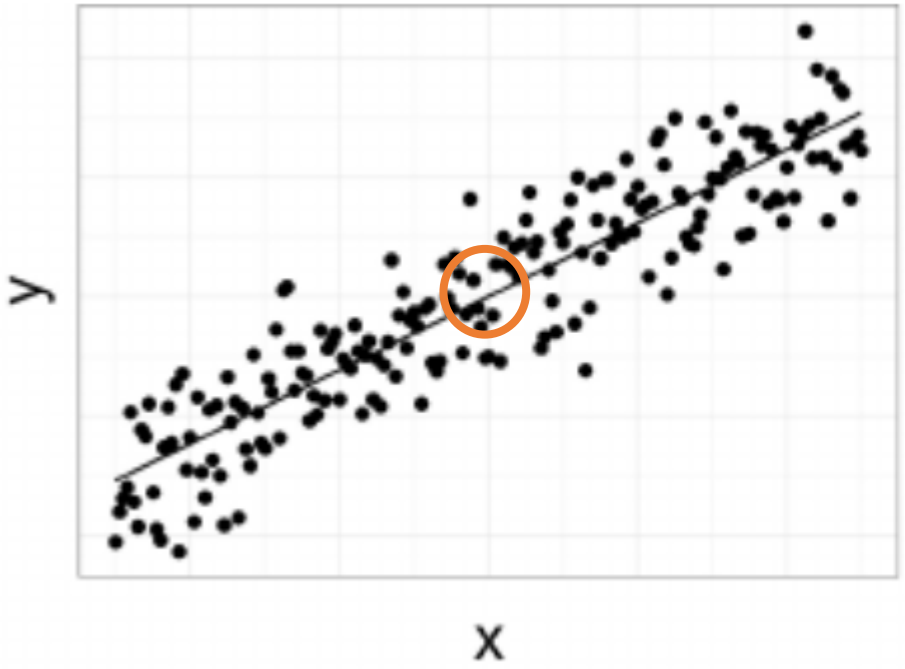
In which example does our regression line best reflect the mean of all the Y values, for any given x?

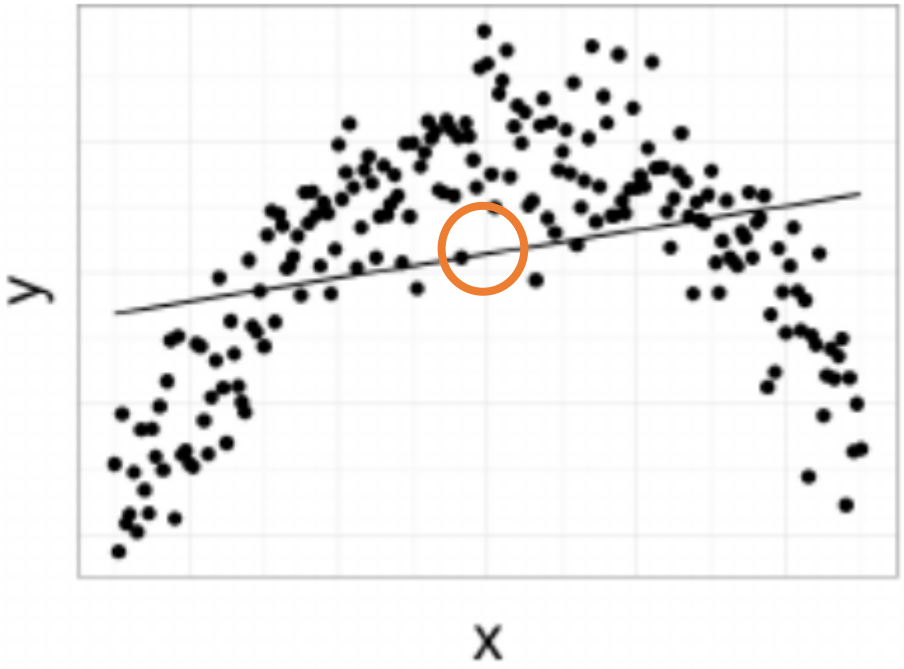




A prediction within the orange circle on the right would probably not reflect the mean of Y at that given X value.

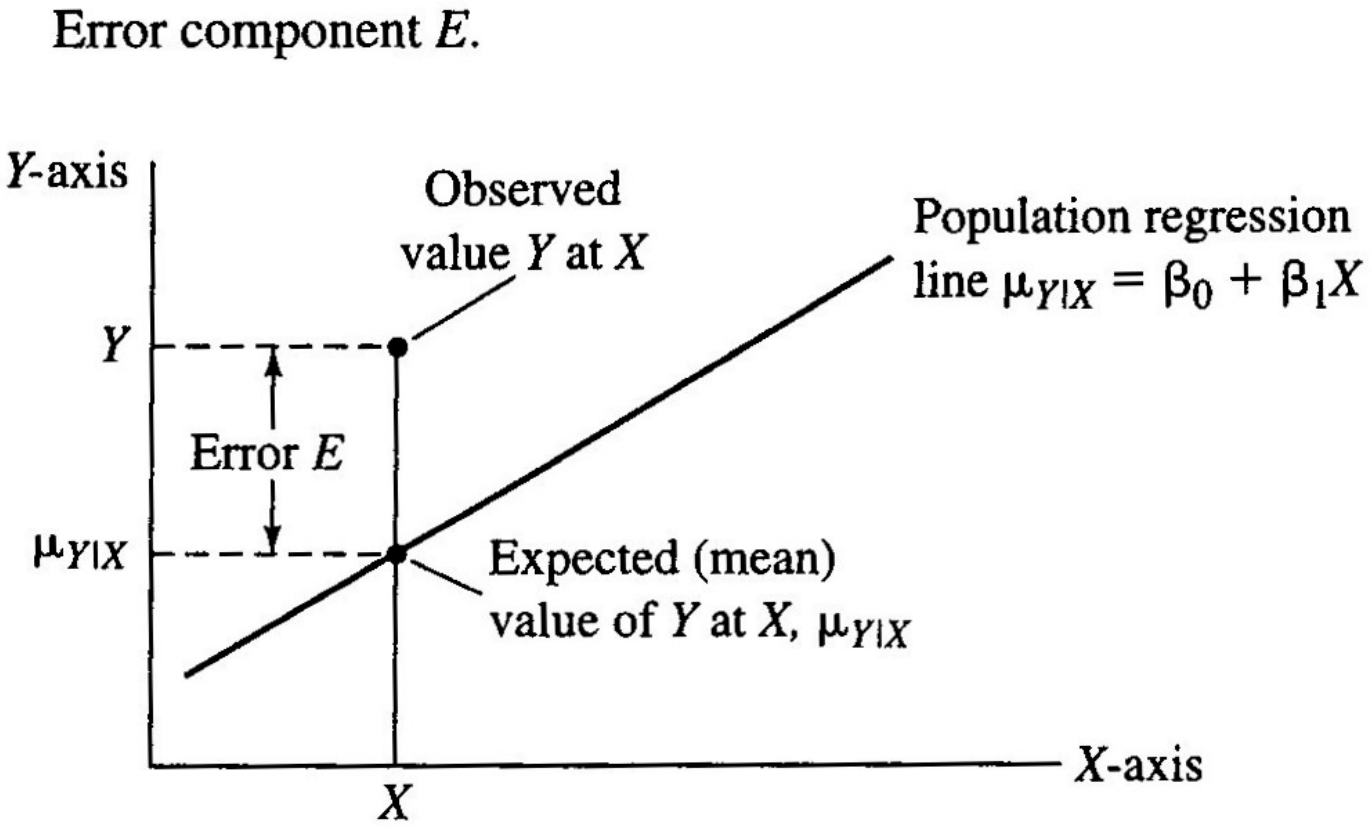
We would likely see that in the model on the right, the residuals (*e*) would be quite large.





Note that a model that fits well will have small residuals.

We'll talk more later about how analyzing the residuals can show us how well the model fits the data.



Independence

Conditional on X, the Y-values should be statistically independent of one another.

That is, the error terms (e) should be independent of one another.

Do you think that the following situations reflect independence of residuals?

- ❑ A chart review of comorbidities of patients who were undergoing COVID treatment at LAC+USC hospital.
- ❑ A study on visual acuity at Roski Eye Institute where individuals' visual acuity was measured in the left and right eyes.
- ❑ The height of children was measured each year for 5 years (longitudinally).
- ❑ The BMI of children and their guardian (mother or father) was assessed in a study on physical activity.
- ❑ A study on the GDP of several countries and the population's collective attitude toward certain policies.

Lack of independence can be accounted for by more complex regression models such as GEE, mixed-effect models, genetic models, etc.

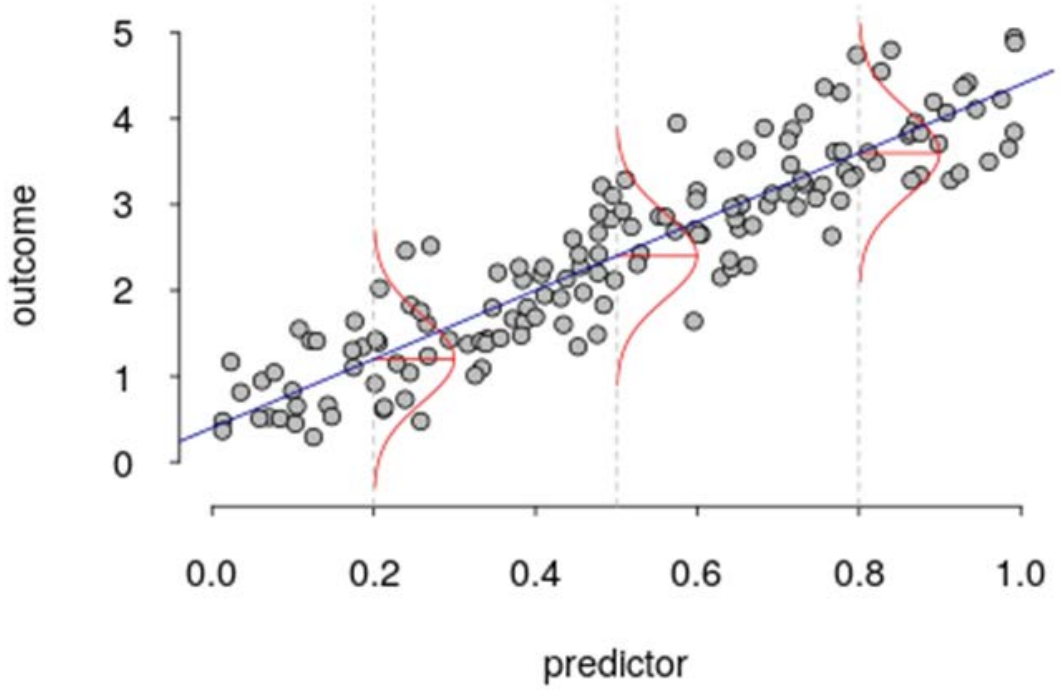
PM511c is a great course to take to learn the skills for analyzing correlated data!

Normality

For any fixed value of X, Y has a normal distribution (i.e., Y|X ~ N).

Equivalently, we can say that the residuals are normally distributed.

Remember the Central Limit Theorem – it makes inferences "robust" to deviations from this assumption when sample size is large.

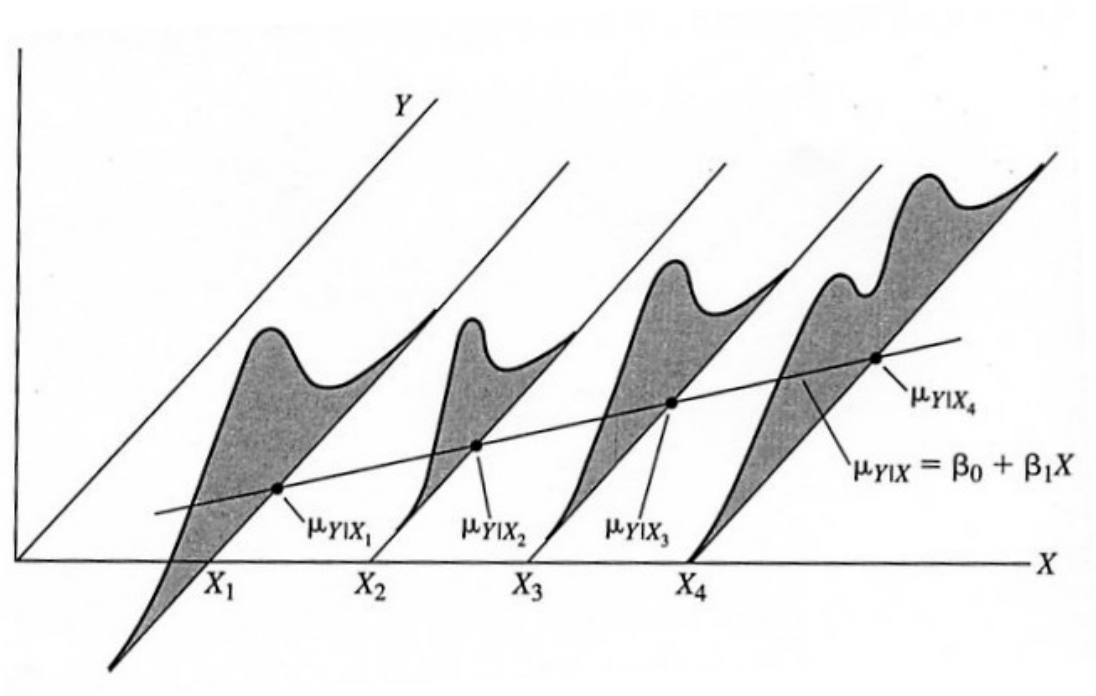


Note that this is different from Y~N! However, if Y is not normally distributed then Y|X may not be normally distributed either.

Equal Variance (Homoscedasticity)

For any value of X, the variance of Y is constant (i.e., $\sigma^2_{Y|X} = \sigma^2$).

In the following figure, the assumptions of normality and homoscedasticity are violated – why?



The normality and homoscedasticity assumptions are equivalent to saying that $e \sim N(0, \sigma^2)$.

Recap

- Linear models are only valid if the “LINE” assumptions hold
- Assessing these assumptions is important when determining the validity of the model

Recap

- State the four LINE assumptions
- Explain how to assess each of the assumptions

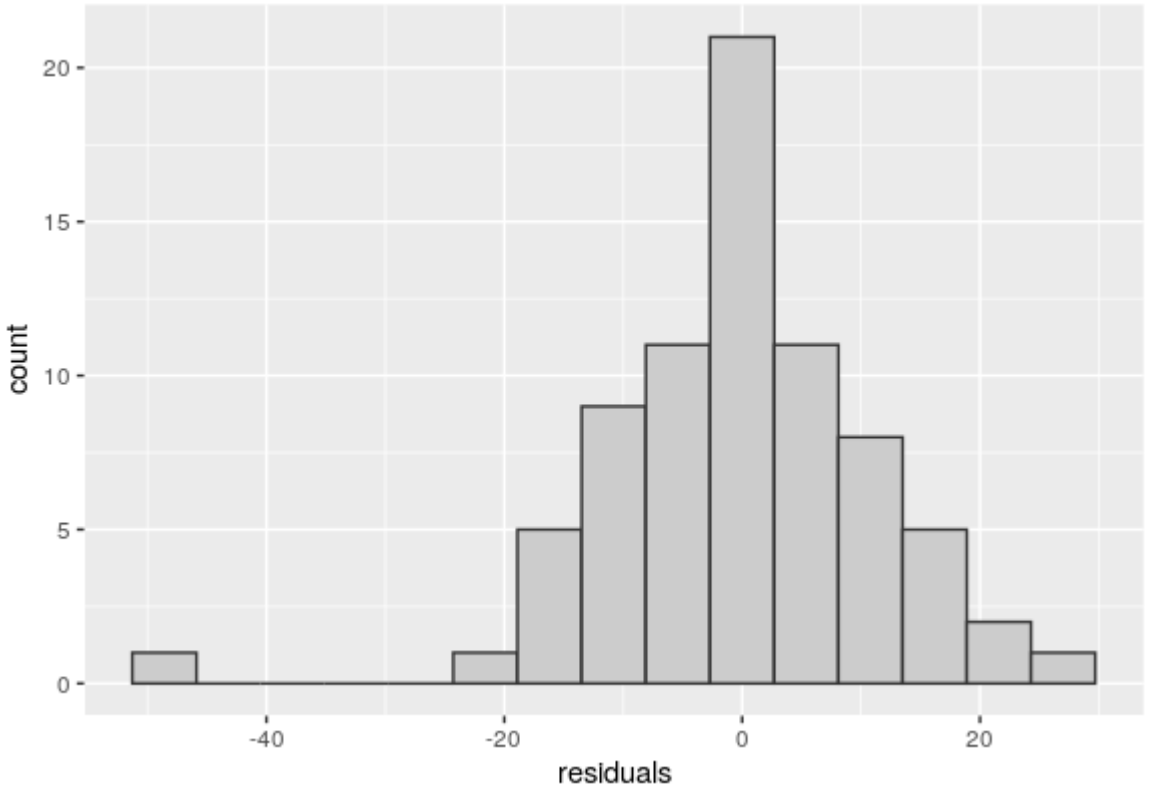
Test Yourself

You examine the residuals from a model you created. These residuals tell us:

a. All observations fit the model well.

b. One observation had a predicted value far lower than the observed value.

c. One observation had a predicted value far higher than the observed value.



Test Yourself

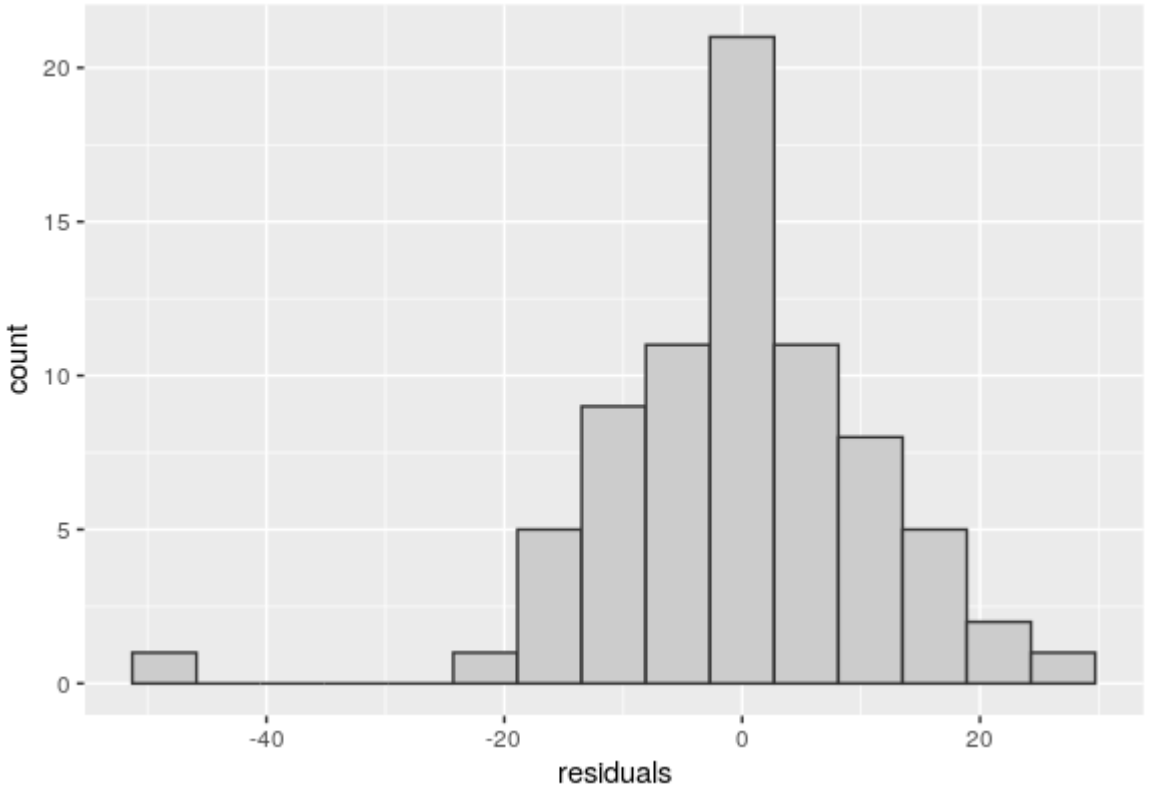
Remember, the residuals equal the observed value minus the expected/predicted value. If Obs-Exp=Resid, and Resid<0, then Obs-Exp<0 and therefore Obs<Exp

You examine the residuals from a model you created. These residuals tell us:

a. All observations fit the model well.

b. One observation had an expected value far lower than the observed value.

c. One observation had an expected value far higher than the observed value.



The Best-Fit Line

We have an intuition about what the best line for Y on X would look like, but how is it actually found? How did R give us the estimates of intercept and slope?

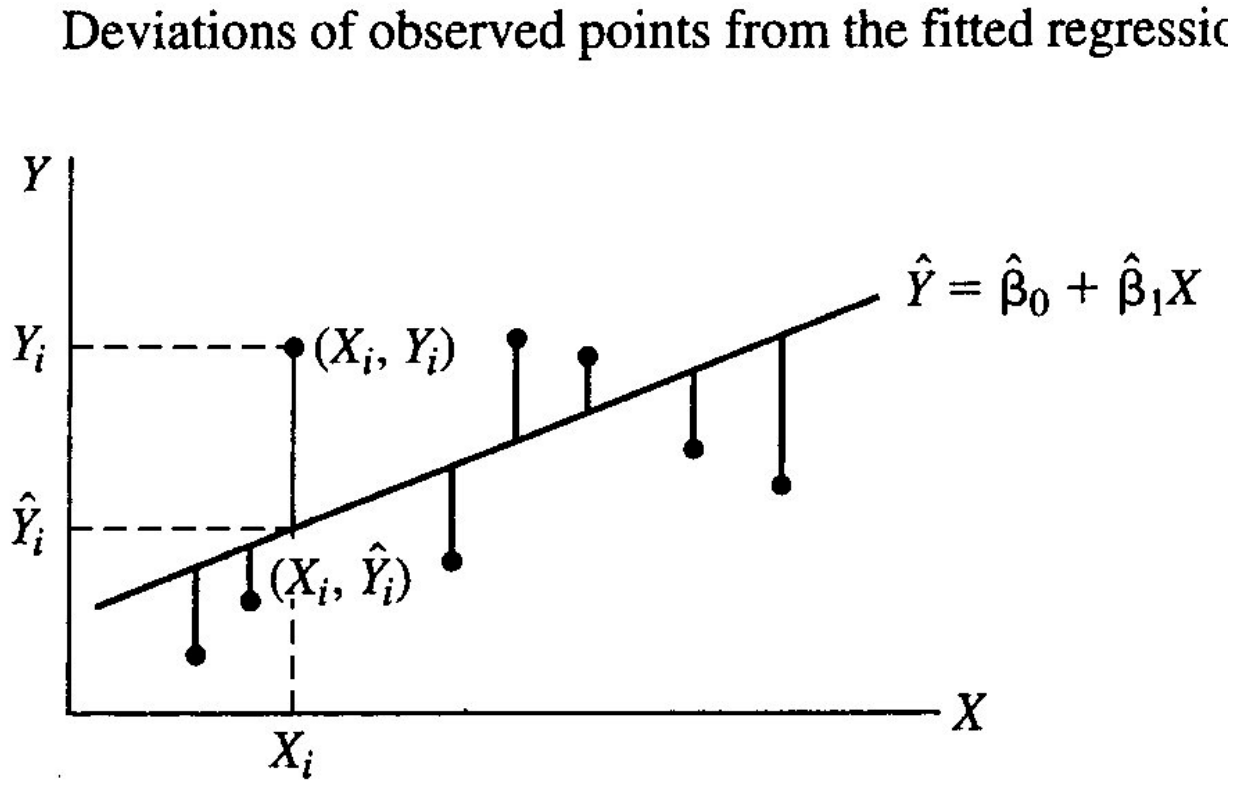
There is *more than one* way to arrive at a result in this sense.

In linear regression these methods will provide the same results.

Option 1: Ordinary Least Squares (OLS)

Find the line that minimizes the sum of squares of the residuals.

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are the predicted values that come from the regression-fitting approach.



The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ are called the least squares estimates if they are arrived at by minimizing the square of the residuals.

This operationally involves minimizing:

$$\sum_{i=1}^{N}\left[Y_i - \left(\hat{\beta}_0 + \hat{\beta}X_i\right)\right]^2 =$$

$$\sum_{i=1}^{N} e_i^2$$

This value is the "residual sum of squares" or the "sum of squares due to error" (SSE).

Option 2: Minimum Variance

Find the best linear unbiased estimators (BLUE) $\hat{\beta}_0$ and $\hat{\beta}_1$ that are the "best" because they have minimum variance of all unbiased estimators.

By the Gauss-Markov theorem, for linear regression, this definition will give rise to the same least-squares estimators.

Option 3: Maximum Likelihood

Find the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ that maximize the likelihood of the observed data occurring (more on this with logistic regression later). These are the same as the least-squares estimators in the linear regression context.

Recap

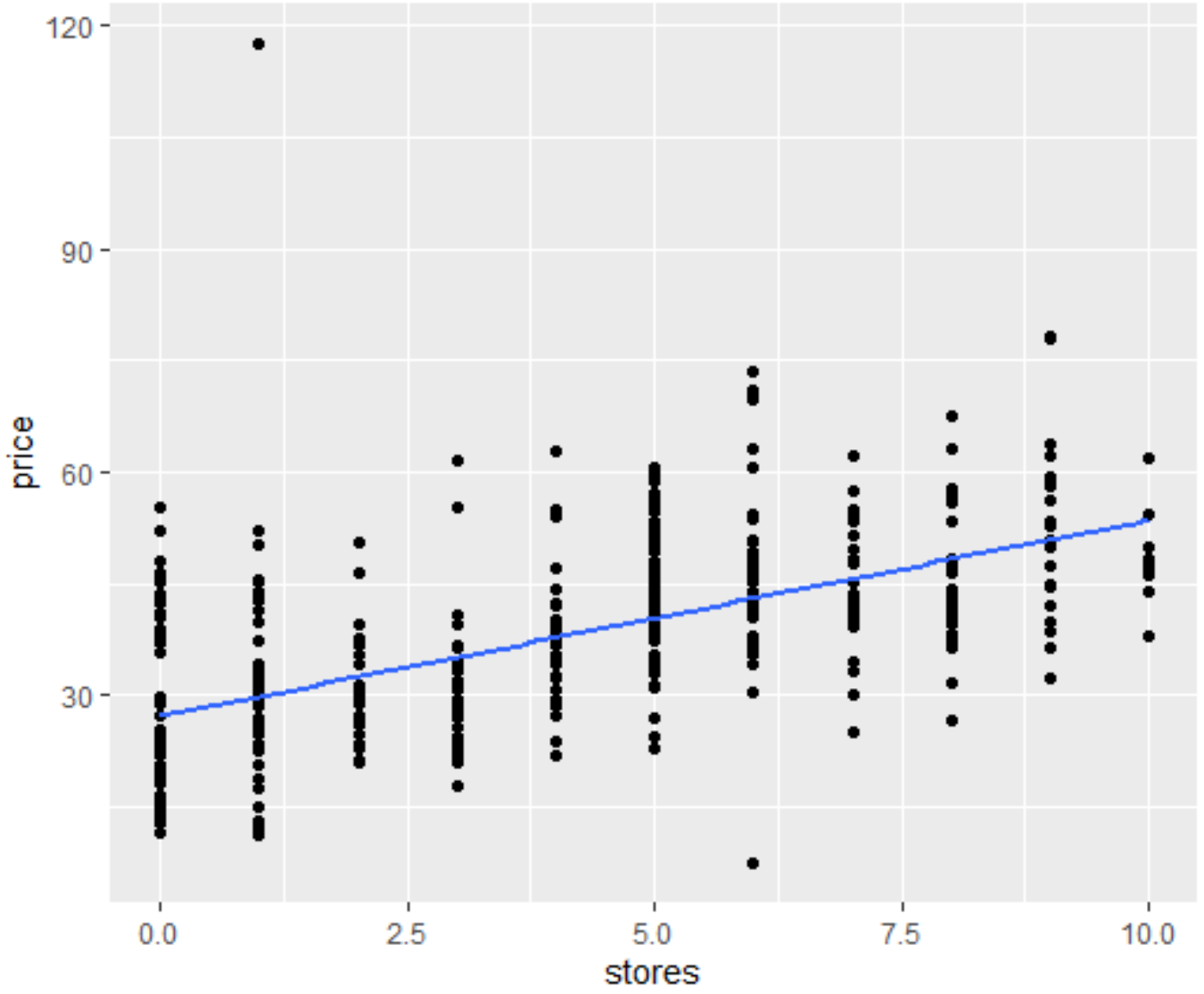
- While software will find the "best" estimates of the model's β coefficients, it is important to know how these estimates are produced
- In the case of simple linear regression, the methods discussed in this section will all produce the same parameter estimates

Recap

- Explain conceptually how the “ordinary least squares” approach arrives at the model’s parameter estimates

The real_estate CSV file contains information on the selling price of homes and other variables that may be associated with this.

Is the number of stores close by related to selling price (per area)?



6. Interpretation of Output

```
> model_price_stores <- lm(price ~ stores, data = realestate)
> summary(model_price_stores)

Call:
lm(formula = price ~ stores, data = realestate)

Residuals:
    Min      1Q  Median      3Q     Max
-35.407  -7.341  -1.788   5.984  87.681

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.1811     0.9419   28.86   <2e-16 ***
stores        2.6377     0.1868   14.12   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.18 on 412 degrees of freedom
Multiple R-squared:  0.326,     Adjusted R-squared:  0.3244
F-statistic: 199.3 on 1 and 412 DF,  p-value: < 2.2e-16

> anova(model_price_stores)
Analysis of Variance Table

Response: price
           Df Sum Sq Mean Sq F value    Pr(>F)
stores      1  24930 24930.0  199.32 < 2.2e-16 ***
Residuals 412  51531   125.1
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This can be useful when examining the residuals. There is one particularly high residual value. See the histogram to the right.

$\hat{Y} = 27.2 + 2.6X_{stores}$

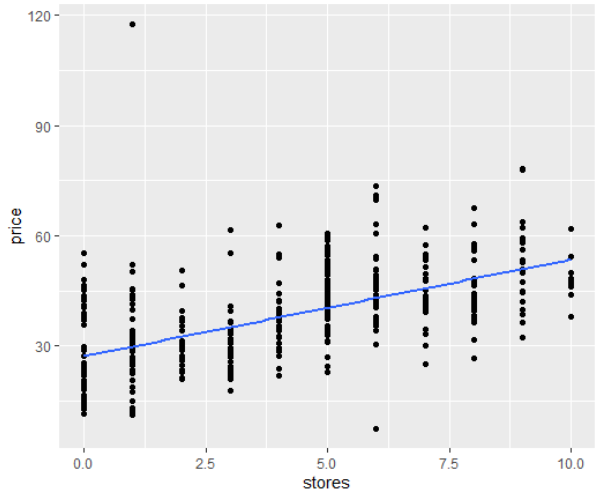
Price is statistically significantly related to # stores (p<.001).

SSE

$s^2_{Y|X}$

$s_{Y|X}$

These values correspond to the part of the model that isn't explained. Better models have lower residual standard error.

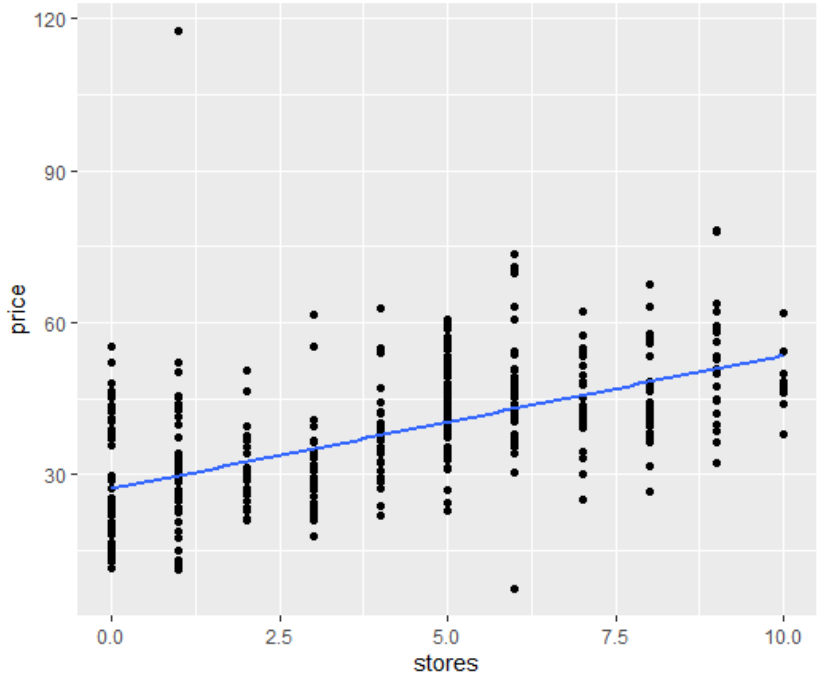


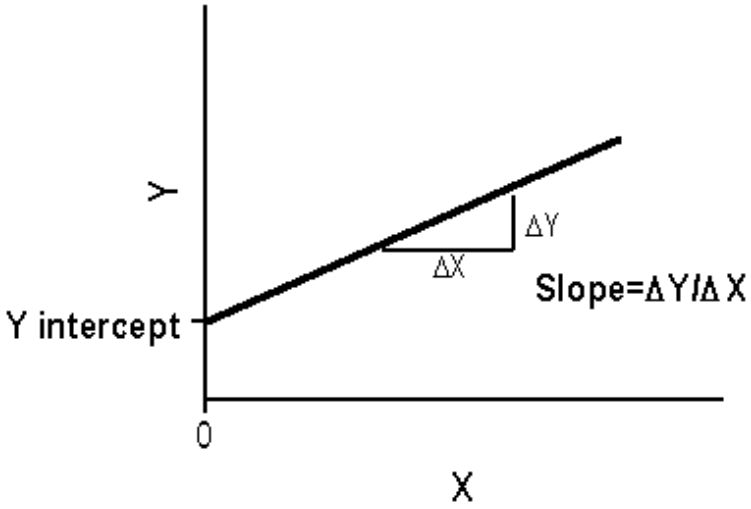
$$\hat{Y} = 27.2 + 2.6X_{stores}$$

Slope. For each additional store that is close to the house, the estimated mean price (i.e., predicted price) increases by 2.6.

Intercept. The estimated mean price for a house with no nearby stores is 27.2.

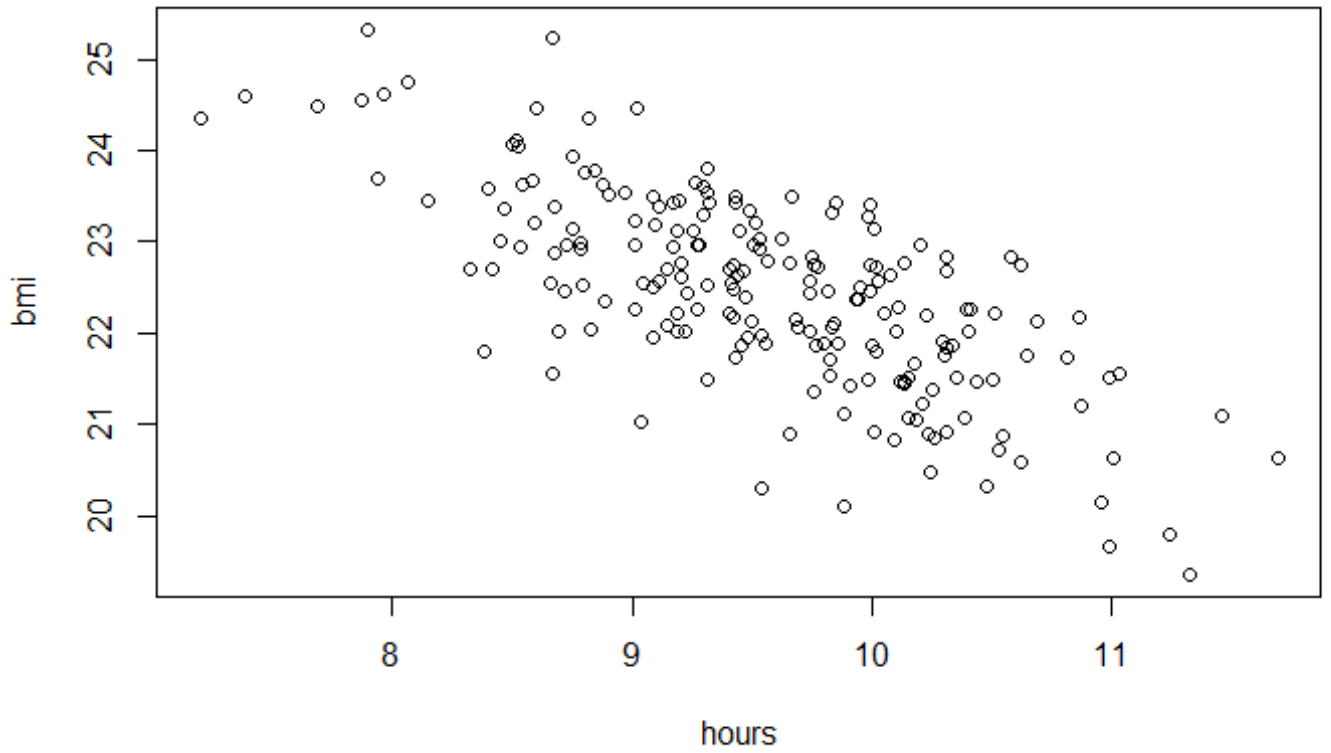
Is the intercept interpretable?





The sleep CSV file contains information on the BMI of adolescents and the daily amount of sleep they get.

Is BMI related to amount of sleep?



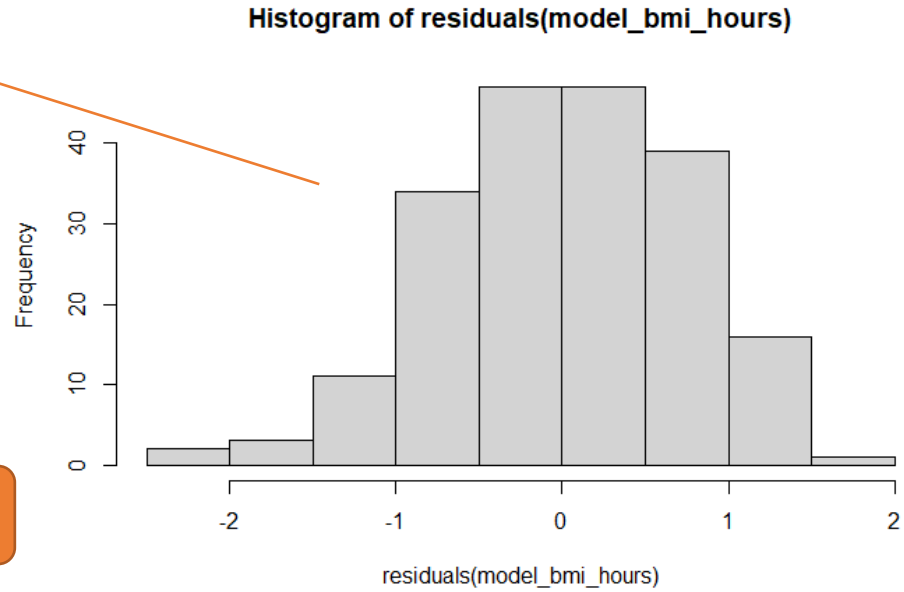
6. Interpretation of Output

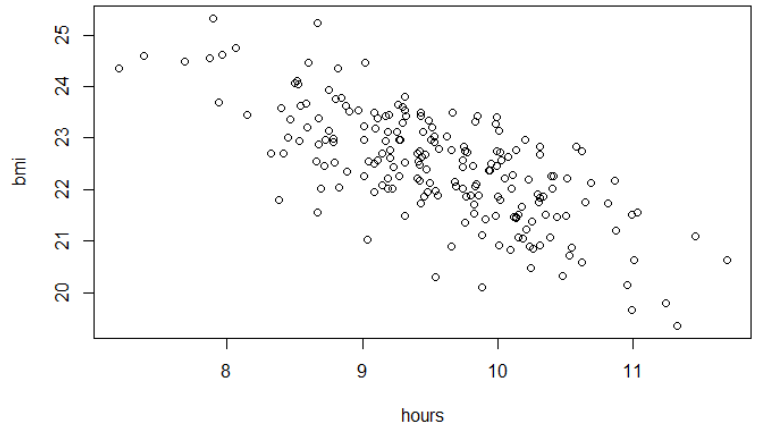

```
> summary(model_bmi_hours)

Call:
lm(formula = bmi ~ hours, data = sleep)

Residuals:
     Min       1Q   Median       3Q      Max
-2.14316 -0.48736  0.02937  0.56833  1.90682

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.77942    0.64391   49.35   <2e-16 ***
hours       -0.97660    0.06715  -14.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0

Residual standard error: 0.7449 on 198 degrees of freedom
Multiple R-squared:  0.5165,     Adjusted R-squared:  0.5141
F-statistic: 211.5 on 1 and 198 DF,  p-value: < 2.2e-16

> anova(model_bmi_hours)
Analysis of Variance Table

Response: bmi
           Df Sum Sq Mean Sq F value    Pr(>F)
hours       1 117.39 117.394  211.54 < 2.2e-16 ***
Residuals 198 109.88   0.555
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The residuals look pretty good! No outliers. Seemingly normal.

$\hat{Y} = 31.78 - 0.98X_{hours}$

Hours is statistically significantly related to BMI (p<.001).

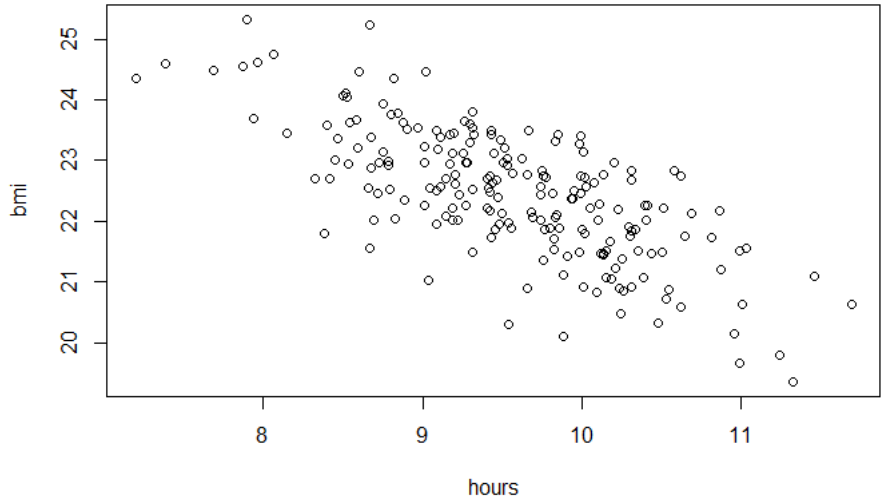




$$\hat{Y} = 31.78 - 0.98X_{hours}$$

Slope. For each additional hour sleep per night that an adolescent gets, BMI is expected to decrease by 0.98.

Intercept. The estimated mean BMI for an adolescent who gets no sleep is 31.78.



Is the intercept interpretable?

To make the intercept more interpretable, it is often desirable to **center the X variable on its mean**.

$$\mu_{Y|X} = \beta_0 + \beta_1(X - \bar{X})$$

```
> summary(model_bmi_hoursc)

Call:
lm(formula = bmi ~ hours_c, data = sleep)

Residuals:
     Min       1Q   Median       3Q      Max
-2.14316 -0.48736  0.02937  0.56833  1.90682

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.44549    0.05268  426.11   <2e-16 ***
hours_c     -0.97660    0.06715  -14.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7449 on 198 degrees of freedom
Multiple R-squared:  0.5165,     Adjusted R-squared:  0.5141
F-statistic: 211.5 on 1 and 198 DF,  p-value: < 2.2e-16
```

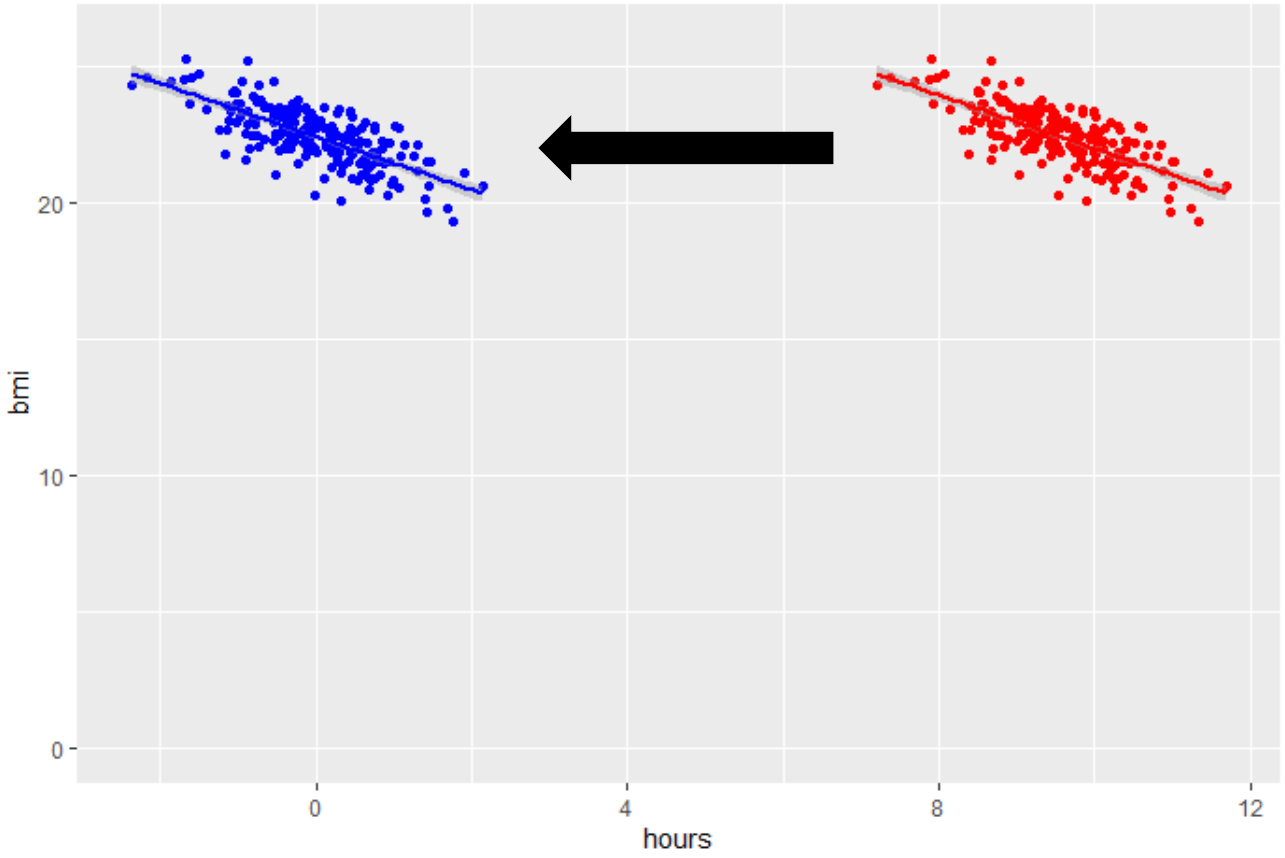
The intercept now has the interpretation: “The mean BMI for individuals who get the mean hours of sleep is 22.45

Centering variables (a linear transformation) will not change the relationship between X and Y.

John Tukey’s Advice:
“Never estimate intercepts – always estimate *centercepts!*”

We took our original data (red) and shifted it to the left (blue).

- The intercept <u>is</u> changed.
- The slope is <u>not</u> changed.
- The SSE is <u>not</u> changed.
- The F-statistic is <u>not</u> changed.



Fun fact: when X is centered on its sample mean, $\hat{\beta}_0 = \bar{Y}$.

Recap

- We have previously learned how to interpret the β estimates from statistical output
- Information about the residuals can tell us how well the model fits
- Centering your X variable can sometimes make the intercept more interpretable, but will not change the fit of the model

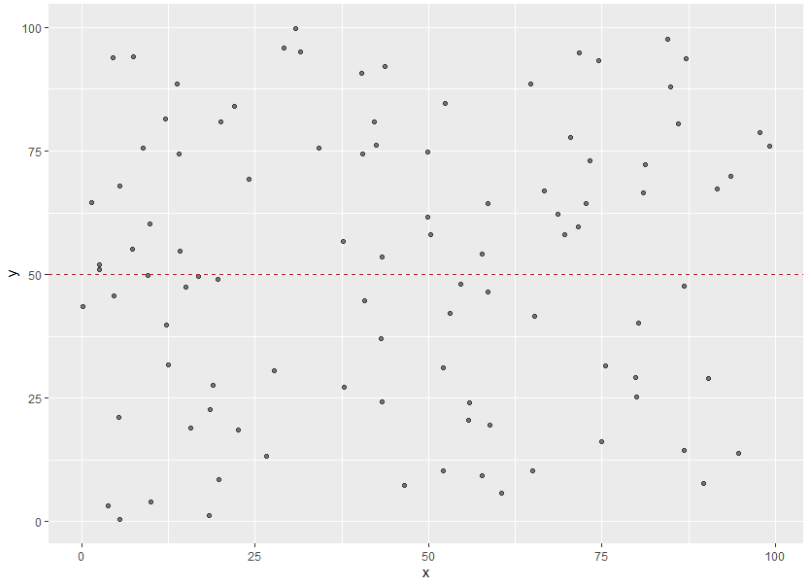
Recap

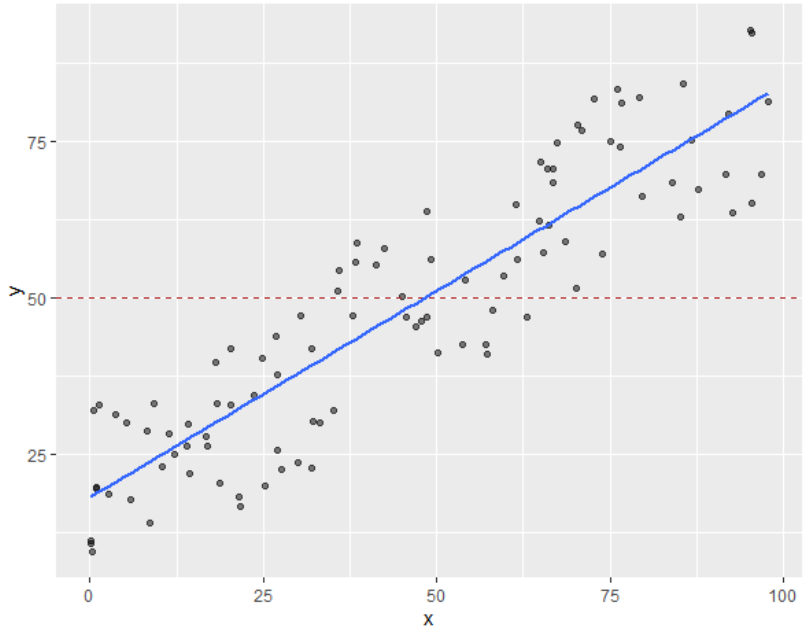
- Explain the meaning of the information about residuals in model output
- Be able to center a variable on its mean (or other value)
- Explain how centering a variable changes the meaning of the intercept

Slope

- H_0: $\beta_1 = 0$ (no correlation/association between X and Y).
- $\mu_{Y|X} = \bar{Y} + \beta_1(X - \bar{X})$, so we can interpret H_0 as saying "our best prediction would be to predict the mean Y value for everyone."
- If H_0 is rejected, then it means that the model under consideration performs better than $\mu_{Y|X} = \bar{Y}$.
- If we fail to reject H_0 then it implies that a flat line (0 slope) does just a good a job as any.

But there could still be a non-linear relationship!





How is the slope tested?

The Wald test:

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \text{ on N-2 df where}$$

$$se(\hat{\beta}_1) = \frac{s_{Y|X}}{s_X\sqrt{n-1}}$$

Standard deviation of the residuals.

Standard deviation of X

Intercept

- H_0: $\beta_0 = 0$ (the true mean of Y when X=0).
- This also follows a Wald test:

$$t = \frac{\hat{\beta}_0}{se(\hat{\beta}_0)}$$

$$se(\hat{\beta}_0) = s_{Y|X}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1)s_X^2}}$$

This is the mean of whatever variable multiplies β. So it could be the mean of X, or the mean of X - $\bar{X}$ (i.e., 0).

Confidence Intervals for Parameters

A (1-α)% confidence interval for any parameter is given by:

$$\hat{\beta} \pm t_{\left(n-2,1-\frac{\alpha}{2}\right)} se(\hat{\beta})$$

Confidence Intervals for The Regression Line

The parameter of interest is $\mu_{Y|X_0}$, the mean value of Y for a given value of X (i.e., $X=X_0$).

And we know the point estimate is given as $\hat{\mu}_{Y|X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$

So the (1-α)% confidence interval for $\mu_{Y|X_0}$ is

$$\hat{\mu}_{Y|X_0} \pm t_{\left(n-2,1-\frac{\alpha}{2}\right)} se(\hat{\mu}_{Y|X_0})$$

$$se(\hat{\mu}_{Y|X_0}) = s_{Y|X}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{(n-1)s_X^2}}$$

Confidence Intervals for an Individual Predicted Value

The parameter of interest is Y_{X_0}, the value of Y for a given value of X (i.e., $X=X_0$).

The point estimate is given as $\hat{Y}_{X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0 + \hat{e}$

The error term should be distributed N(0, 1), so our best guess of the error is 0.

Therefore this point estimate is equivalent to $\hat{\mu}_{Y|X_0}$.

The (1-α)% confidence interval for $\hat{Y}_{X_0}$ is

$$\hat{Y}_{X_0} \pm t_{\left(n-2,1-\frac{\alpha}{2}\right)} se(\hat{Y}_{X_0})$$

$$se(\hat{Y}_{X_0}) = s_{Y|X}\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{(n-1)s_X^2}}$$

This value is larger than $\text{se}(\hat{\mu}_{Y|X_0})$.

Conceptually this makes sense; there will be more uncertainty when predicting a specific point vs. predicting the mean.

Let's obtain confidence intervals for our model parameters.

```
> summary(model_bmi_hoursc)

Call:
lm(formula = bmi ~ hours_c, data = sleep)

Residuals:
     Min       1Q   Median       3Q      Max
-2.14316 -0.48736  0.02937  0.56833  1.90682

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.44549    0.05268  426.11   <2e-16 ***
hours_c     -0.97660    0.06715  -14.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7449 on 198 degrees of freedom
Multiple R-squared:  0.5165,     Adjusted R-squared:  0.5141
F-statistic: 211.5 on 1 and 198 DF,  p-value: < 2.2e-16

> confint(model_bmi_hoursc)
                2.5 %     97.5 %
(Intercept) 22.34161 22.5493636
hours_c     -1.10901 -0.8441847
```

$$\hat{\beta}_0 = 22.45\ (95\%CI = 22.34, 22.55)$$
$$\hat{\beta}_1 = -0.98\ (95\%CI = -1.11, -0.84)$$

We can also obtain predicted values for each data point, with confidence intervals.

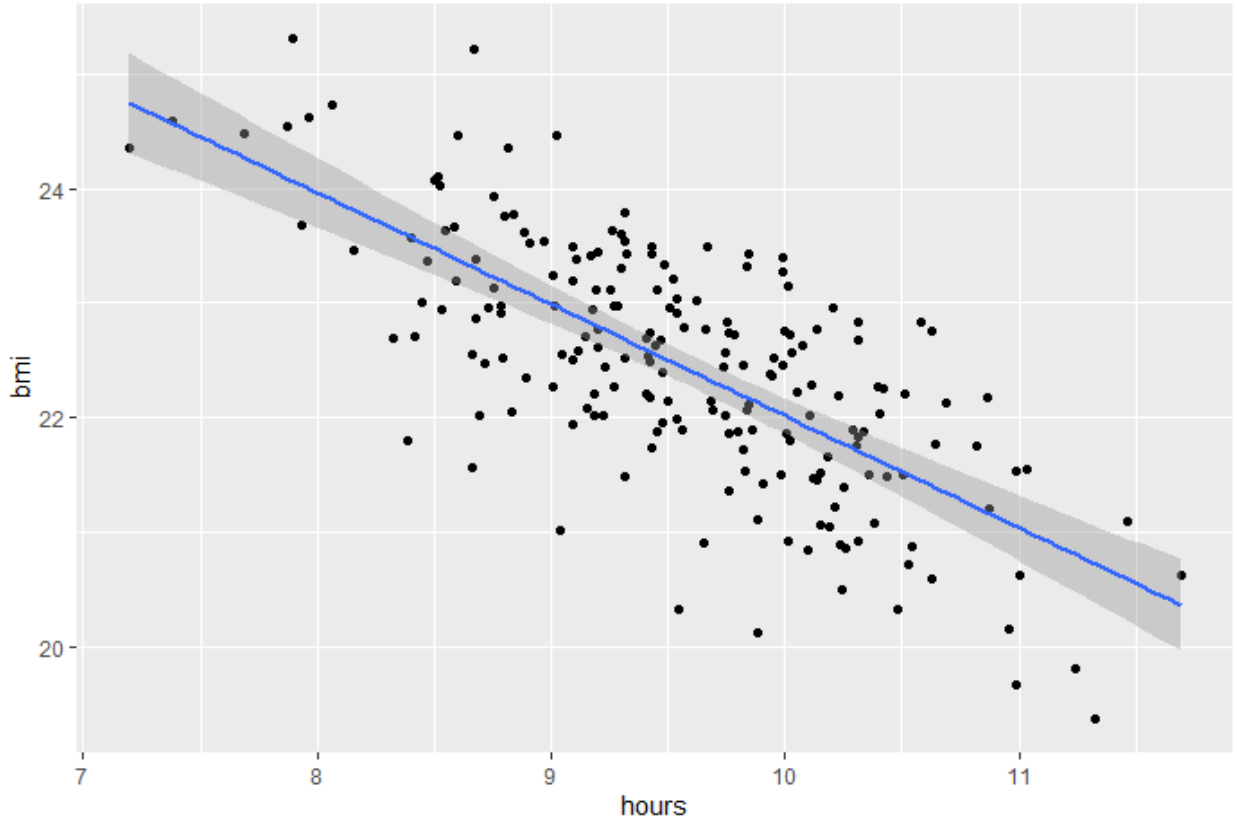
Typically, “confidence interval” is used for the line (mean), while a “prediction interval” is used for the predicted values (points).

```
> predict(model_bmi_hoursc, sleep, interval="confidence") %>%
+   head()
       fit      lwr      upr
1 22.18878 22.07922 22.29833
2 21.99701 21.87665 22.11738
3 23.31543 23.15826 23.47260
4 21.44966 21.27931 21.62002
5 22.68217 22.57345 22.79090
6 22.34976 22.24507 22.45444

> predict(model_bmi_hoursc, sleep, interval="prediction") %>%
+   head()
       fit      lwr      upr
1 22.18878 20.71565 23.66191
2 21.99701 20.52304 23.47099
3 23.31543 21.83799 24.79286
4 21.44966 19.97077 22.92856
5 22.68217 21.20910 24.15524
6 22.34976 20.87698 23.82254
```

When looking at the first 6 observations, the prediction interval is larger than the confidence interval.

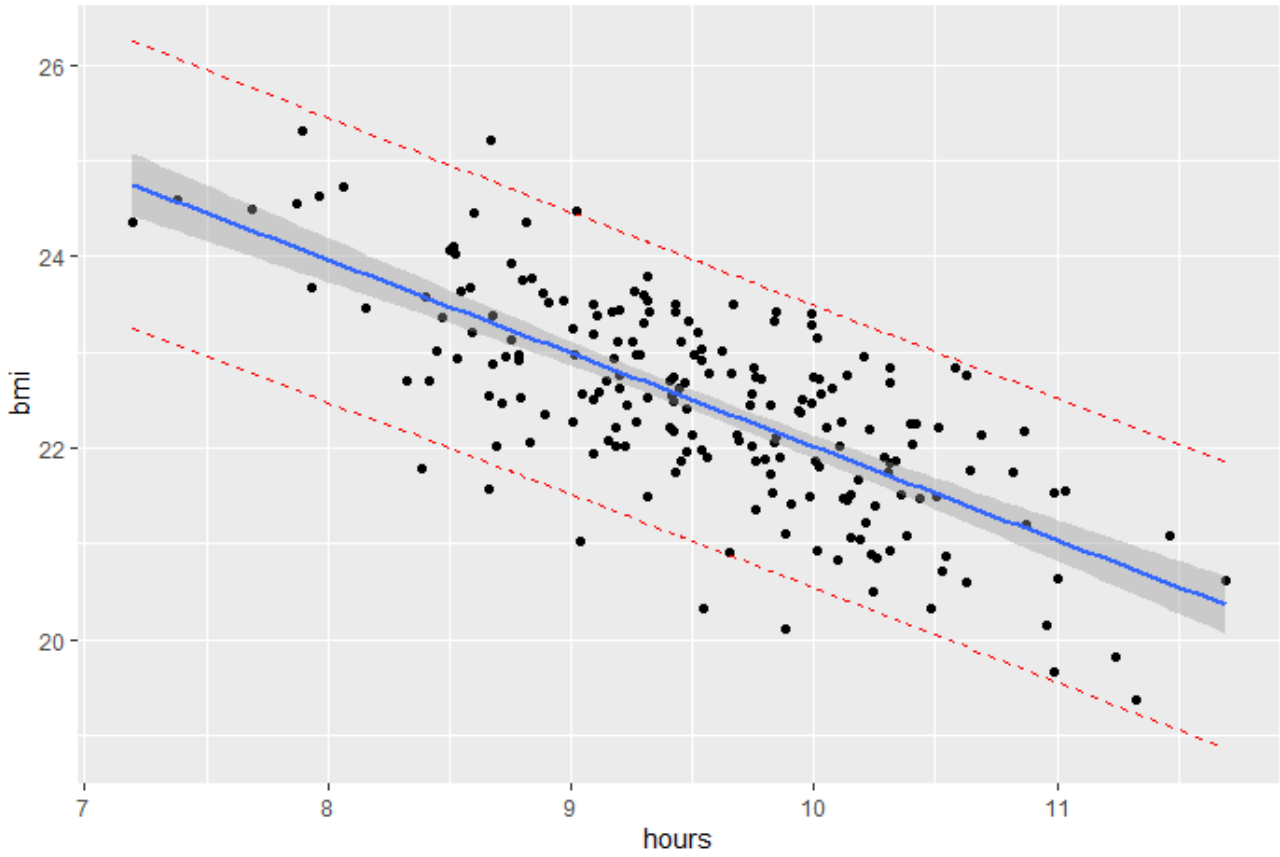
ggplot, by default, includes the 95% confidence intervals when adding a prediction line. This option can be turned off with the "se=" option.

```
sleep %>%
  ggplot(aes(x = hours, y = bmi)) +
  geom_point() +
  geom_smooth(method = "lm", level = 0.95)
```



The 95% prediction intervals are a little more difficult to get, but still do-able.

```
sleep2 <- cbind(
  sleep,
  predict(model_bmi_hoursc, sleep, interval="prediction")
)

sleep2 %>%
  ggplot(aes(x = hours, y = bmi))+
  geom_point() +
  geom_line(aes(y=lwr), color = "red", linetype = "dashed")+
  geom_line(aes(y=upr), color = "red", linetype = "dashed")+
  geom_smooth(method=lm, se=TRUE)
```



Recap

- Without knowing any information about predictors, the best estimate of Y ($\hat{Y}$) is the mean of Y ($\bar{Y}$)
- The Wald test is used to test the hypotheses about slopes:

$$H_0: \beta = 0$$
$$H_A: \beta \neq 0$$

- A 95% confidence interval can be computed on:
  - The regression line, which is the predicted mean of Y across the range of X ($\mu_{Y|X_0}$)
  - The individual predicted values of Y across the range of X ($\hat{Y}_{X_0}$)

Recap

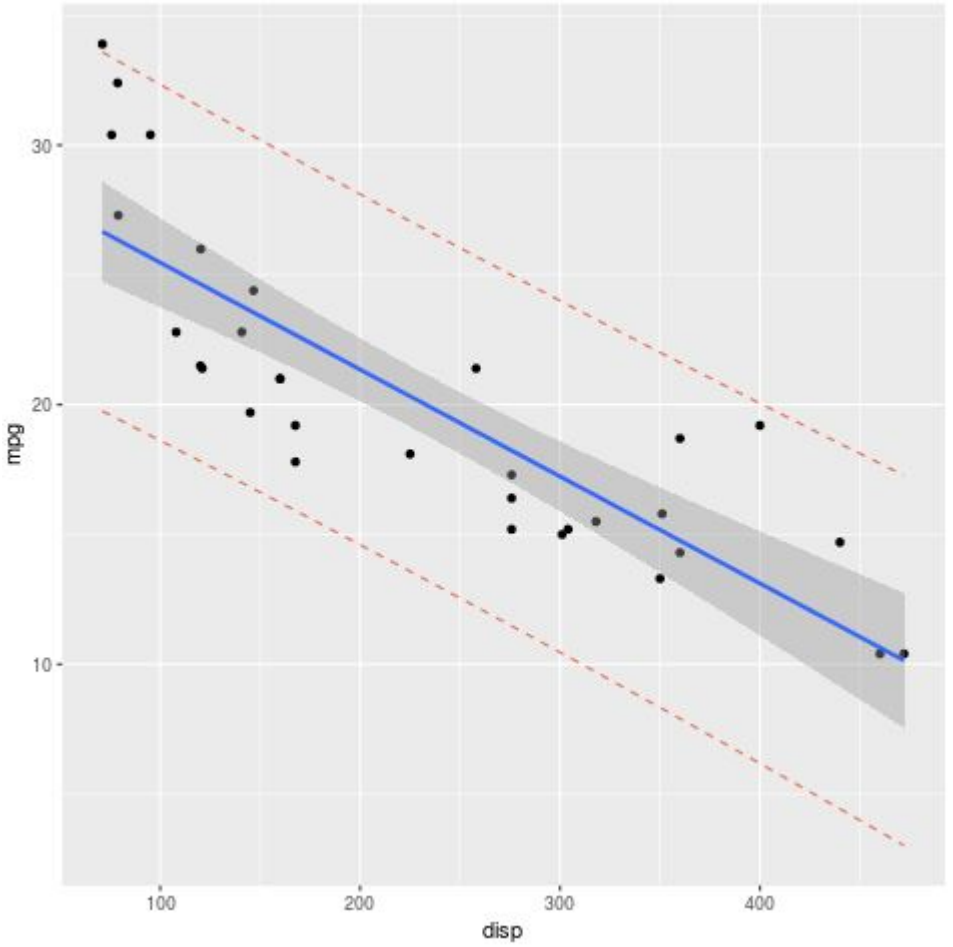
- Describe conceptually the meaning of a test of $\beta = 0$
- Explain and compute the Wald test for β
- Explain and compute the confidence interval for $\mu_{Y|X_0}$
- Explain and compute the prediction interval for $\hat{Y}_{X_0}$
- Plot the corresponding confidence and prediction intervals in R

Test Yourself

Consider the following regression. If a new value was sampled from the same population and added to the data set, we would be 95% confident it would be within the:

a. Gray shaded confidence interval region.

b. Orange dashed prediction interval region.

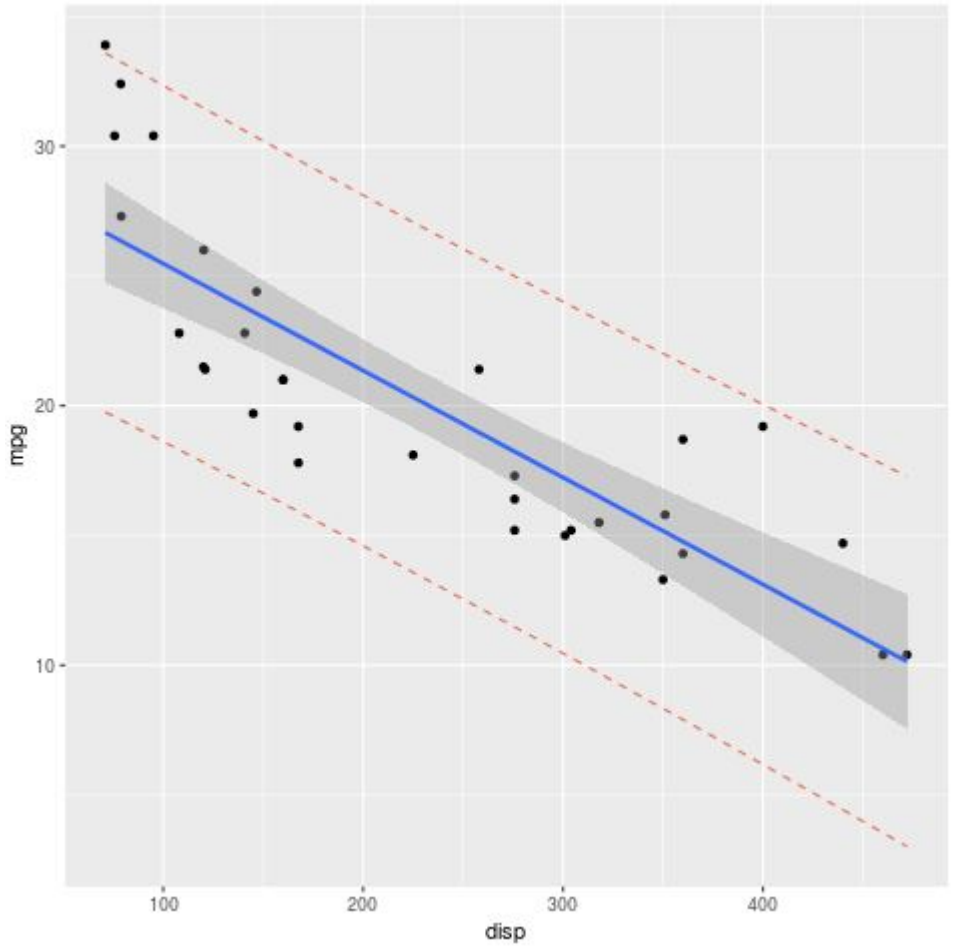


Test Yourself

Consider the following regression. If a new value was sampled from the same population and added to the data set, we would be 95% confident it would be within the:

a. Gray shaded confidence interval region.

b. Orange dashed prediction interval region.



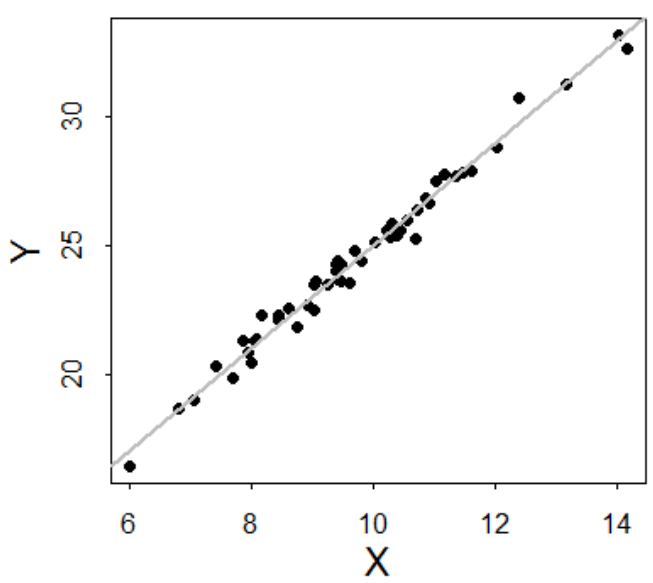
Pearson's Correlation Coefficient

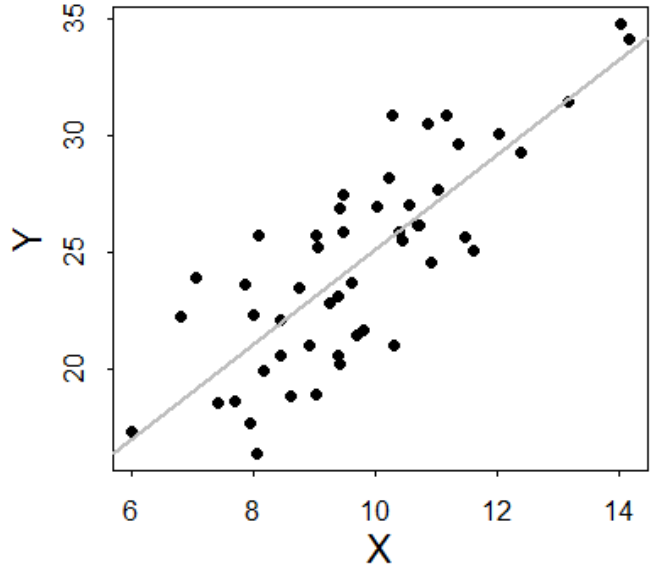
What do the following pieces of information tell us?

Correlation Coefficient: how strongly two variables are related

Regression Coefficients: the nature of how two variables are related.

Take, for example, the following associations.





Both figures show similar least-square regression lines.

However the regression seems to “fit” better in the left figure.

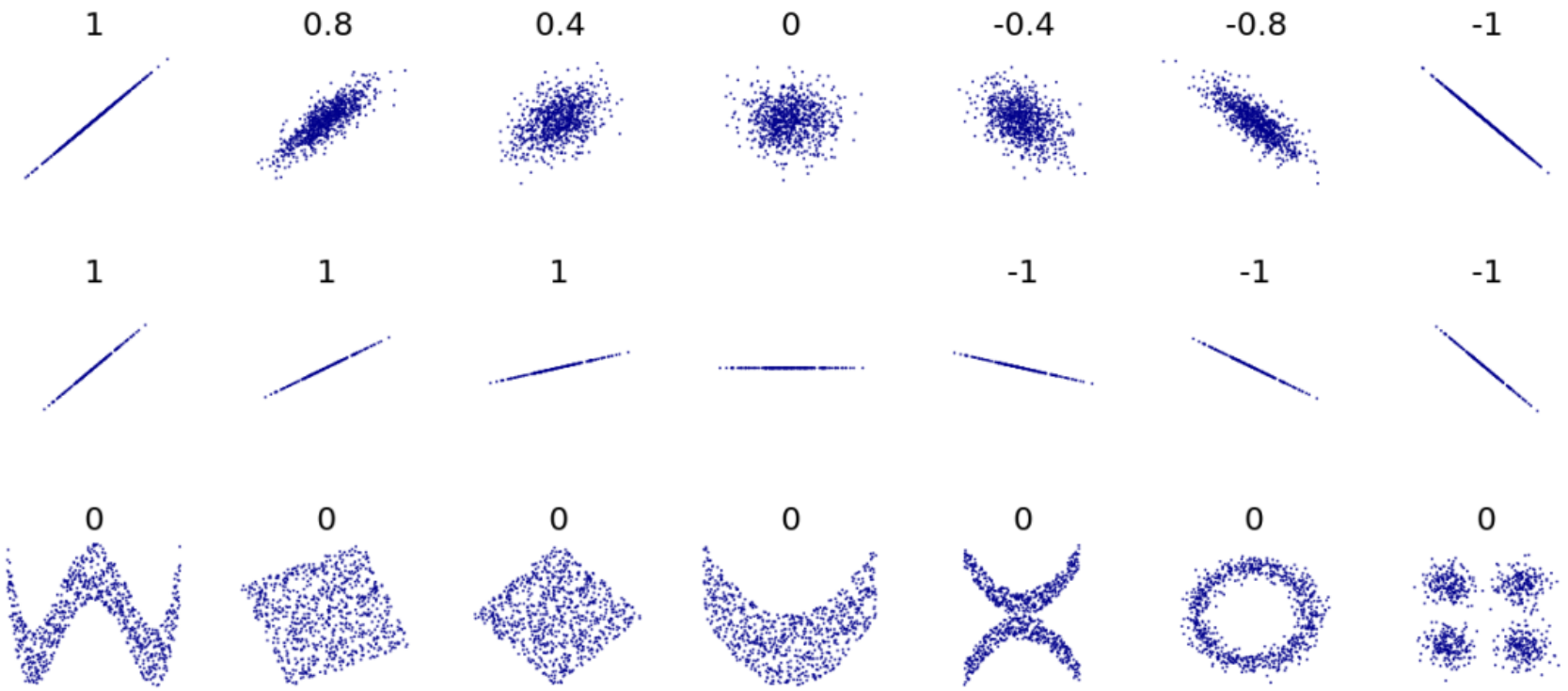
Pearson's correlation coefficient *R* tells us the strength of which two variables are related.

This measure generally has a standard meaning (i.e., it has no units), and ranges from:

- R=-1: perfect negative correlation
- R=1: perfect positive correlation
- R=0: no relationship

Value	Effect Size
0	None
±0.1	Small
±0.3	Medium
±0.5	Large

Just like linear regression, Pearson's correlation can **only detect linear relationships**. For example, none of the relationships in row 3 below would be detected by Pearson's correlation.



The formula for R

$$R = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \frac{Cov(X,Y)}{s_X s_Y}$$

Hypothesis testing for ρ, the population correlation

$H_0: \rho = 0$

$H_A: \rho \neq 0$

$$t = \frac{R}{se(R)} = \frac{R\sqrt{N-2}}{\sqrt{1-R^2}}$$

And has a t-distribution with N-2 df.

For example, we see a strong relationship between hours sleep and BMI in adolescents.

```
> cor(sleep)
              hours         bmi     hours_c
hours     1.0000000 -0.7187008  1.0000000
bmi      -0.7187008  1.0000000 -0.7187008
hours_c   1.0000000 -0.7187008  1.0000000

> with(sleep,
+       cor.test(hours, bmi)
+       )

            Pearson's product-moment correlation

data:  hours and bmi
t = -14.544, df = 198, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7796958 -0.6441946
sample estimates:
       cor
-0.7187008
```

The correlation between hours and BMI is -0.72.

Note: this correlation is the same as that between hours (centered) and BMI.

The correlation is statistically significant ($p < .001$).

Note: the hypothesis tests for correlation and regression are equivalent!
$H_0: \rho = 0$ (The population correlation coefficient = 0)
$H_0: \beta = 0$ (The population regression slope = 0)

And the two measures are closely related:

$$R = \frac{s_x}{s_y}\hat{\beta}$$

Note: the hypothesis tests for correlation and regression are equivalent!

$H_0: \rho = 0$ (The population correlation coefficient = 0)

$H_0: \beta = 0$ (The population regression slope = 0)

And the two measures are closely related:

$$R = \frac{S_x}{S_y}\hat{\beta}$$

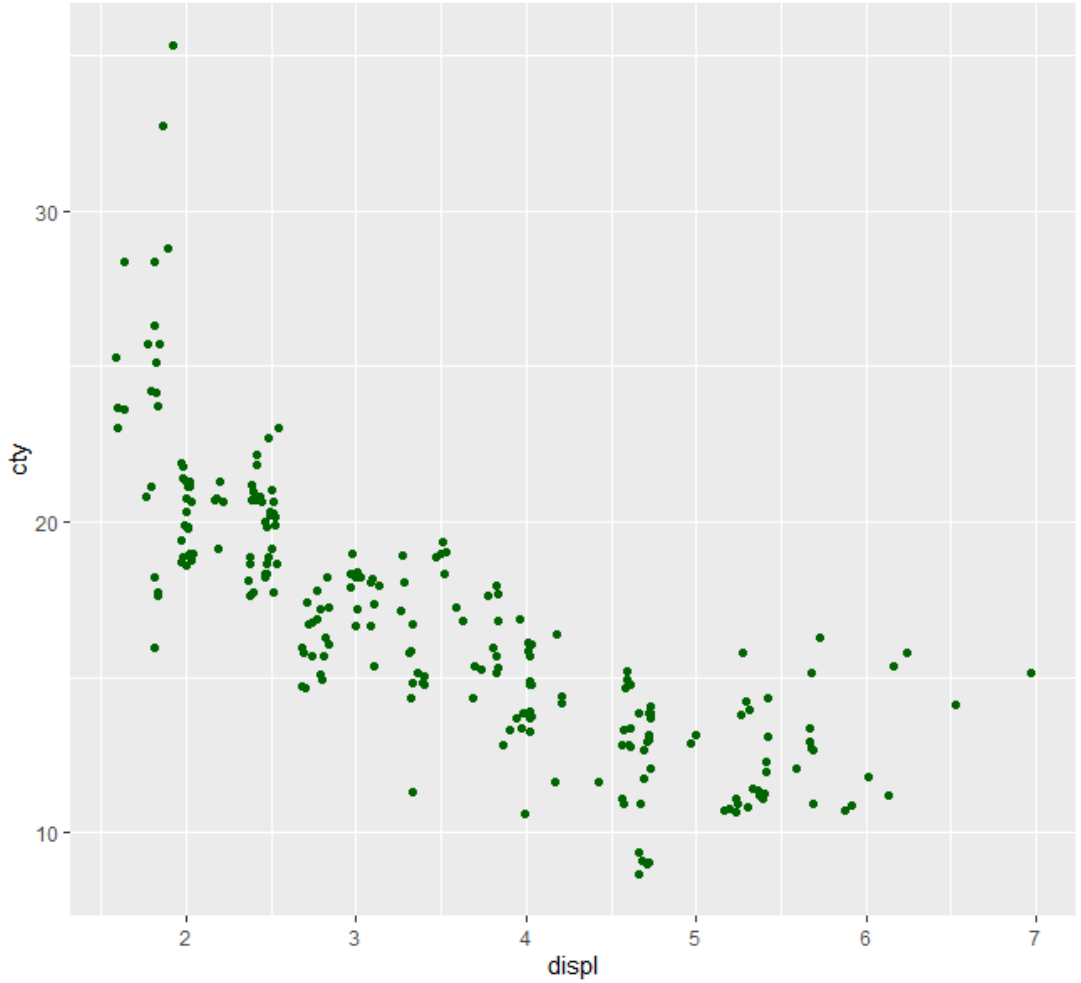
```
> sleep %>%
+   psych::describe()
        vars   n  mean   sd median trimmed  mad   min   max
hours      1 200  9.56 0.79   9.53    9.56 0.77  7.20 11.69
bmi        2 200 22.45 1.07  22.51   22.46 1.01 19.37 25.32
hours_c    3 200  0.00 0.79  -0.02    0.01 0.77 -2.36  2.14

> with(sleep, cor.test(hours, bmi))

data:  hours and bmi
t = -14.544, df = 198, p-value < 2.2e-16
sample estimates:
       cor
-0.7187008
```

```
> summary(model_bmi_hoursc)

Call:
lm(formula = bmi ~ hours_c, data = sleep)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22.44549    0.05268  426.11   <2e-16 ***
hours_c     -0.97660    0.06715  -14.54   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7449 on 198 degrees of freedom
Multiple R-squared:  0.5165,     Adjusted R-squared:  0.5141
F-statistic: 211.5 on 1 and 198 DF,  p-value: < 2.2e-16
```

Let's look at another example – how does a car's engine displacement relate to its fuel efficiency (mpg)?

Do you trust that the correlation coefficient is correct?

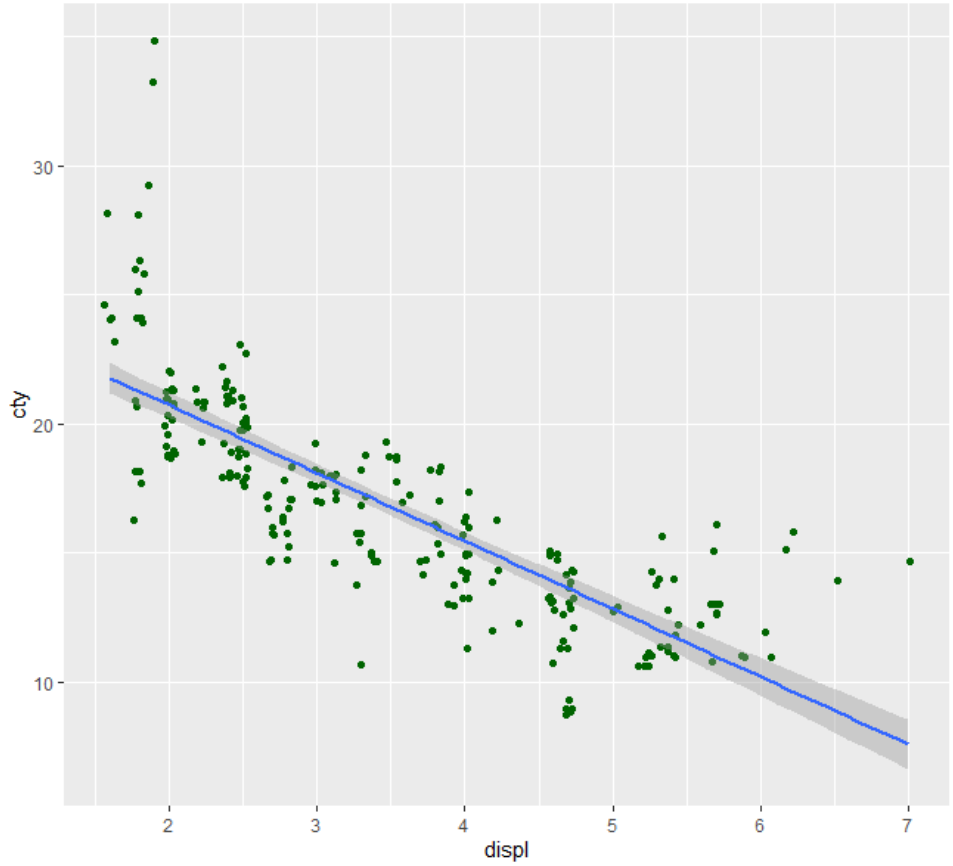
```
> with(mpg,
+      cor.test(cty, displ))

           Pearson's product-moment correlation

data:  cty and displ
t = -20.205, df = 232, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8406782 -0.7467508
sample estimates:
      cor
-0.798524
```



The relationship isn't linear!

What can we do?

1. Apply a transformation to X or Y.
2. Use a nonparametric test.

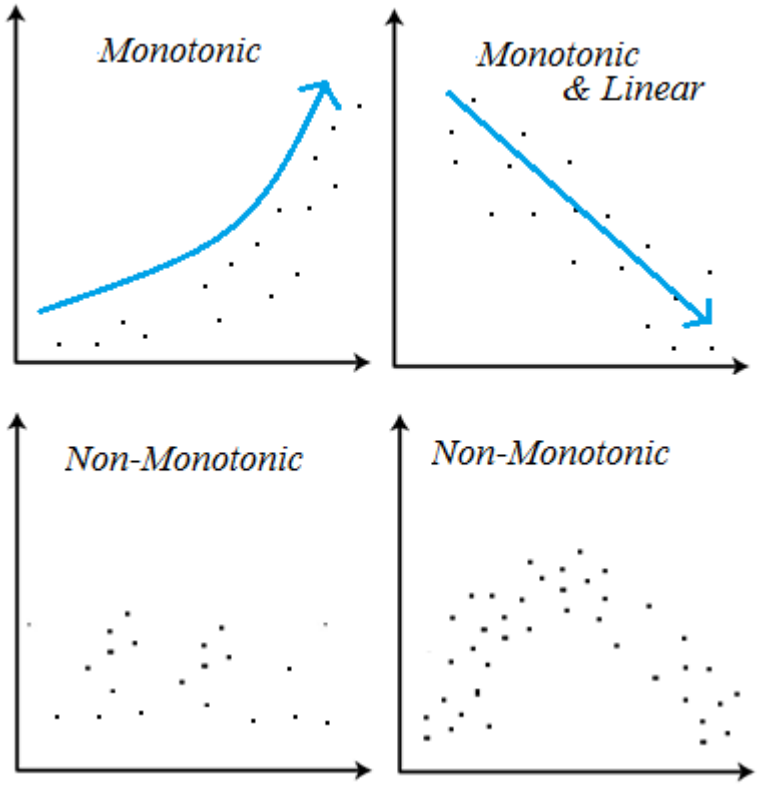


Spearman's R (R_S)

- Used to detect a monotonic relationship between X and Y.
- If a relationship is truly linear, then R will have more power than R_S.
- R_S is equivalent to performing R on the ranks of X and Y.

So if we rank X and Y from low to high, and $d = R_X - R_Y$, then:

$$R_S = 1 - \frac{6\sum_{i=1}^{N} d_i^{\,2}}{N^3 - N}$$



```
> with(mpg,
+      cor.test(cty, displ))

        Pearson's product-moment correlation

data:  cty and displ
t = -20.205, df = 232, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8406782 -0.7467508
sample estimates:
      cor
-0.798524


> with(mpg,
+      cor.test(cty, displ, method = "spearman"))

        Spearman's rank correlation rho

data:  cty and displ
S = 4016569, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.8809049

Warning message:
In cor.test.default(cty, displ, method = "spearman") :
  Cannot compute exact p-value with ties
```

Because the relationship is monotonic, yet nonlinear, the value of R_S=-0.88 is higher in magnitude than R=-0.80.

Recap

- Correlation describes the strength of linear relationship between X and Y, while regression defines the relationship between X and Y
- In simple regression, the results of hypothesis tests for correlation will be the same as those for regression
- Spearman's correlation is a nonparametric version of Pearson's correlation (and is functionally equivalent to running a regression of the rank of Y on the rank of X)

Recap

- Explain the difference between correlation and regression
- Interpret, compare, and contrast regression slope estimates vs. correlation coefficients
- Interpret the results of both Pearson's and Spearman's correlation

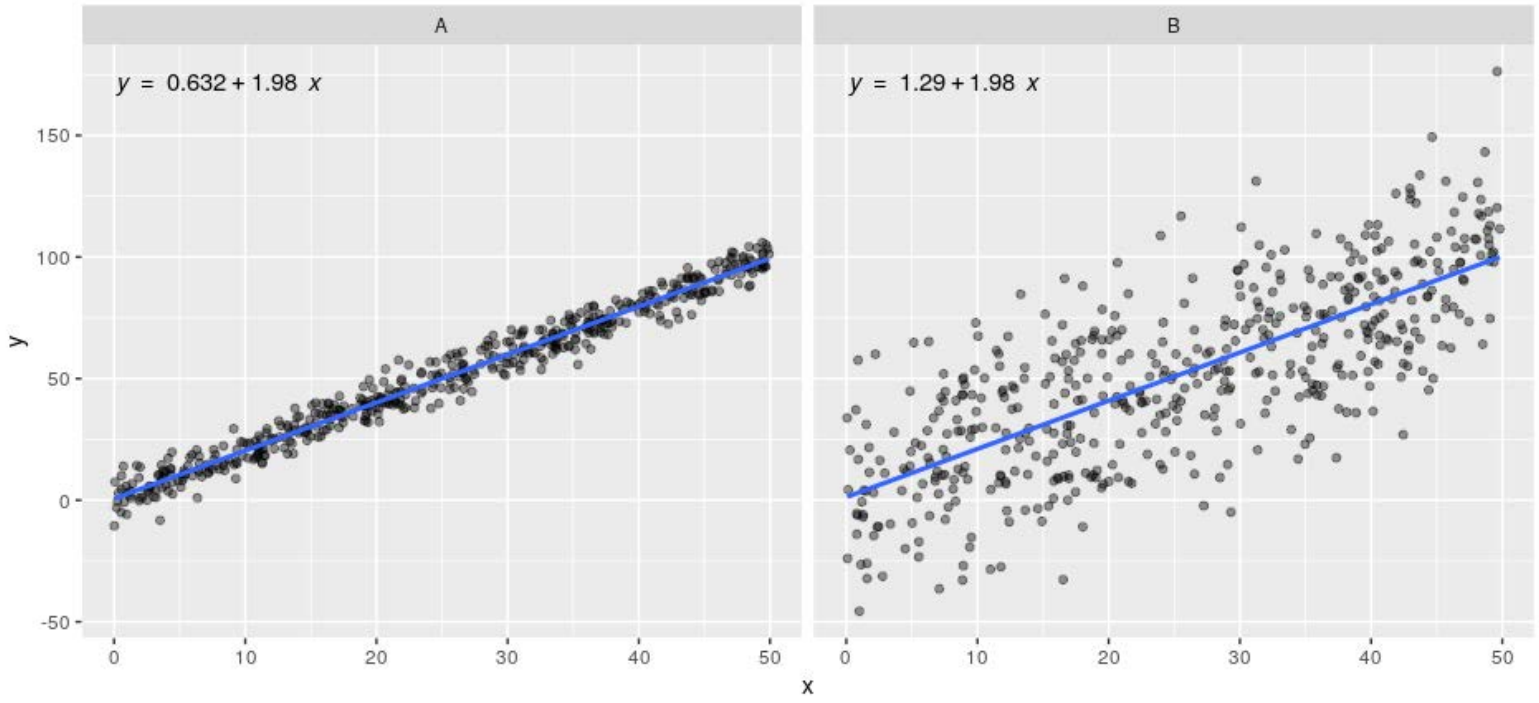
Test Yourself

Consider the regression of Y on X for two groups: A and B.

a. The slope of the regression relationship is larger for group A.

b. The slope of the regression relationship is larger for group B.

c. The slope of the regression relationship is the same between the two groups.



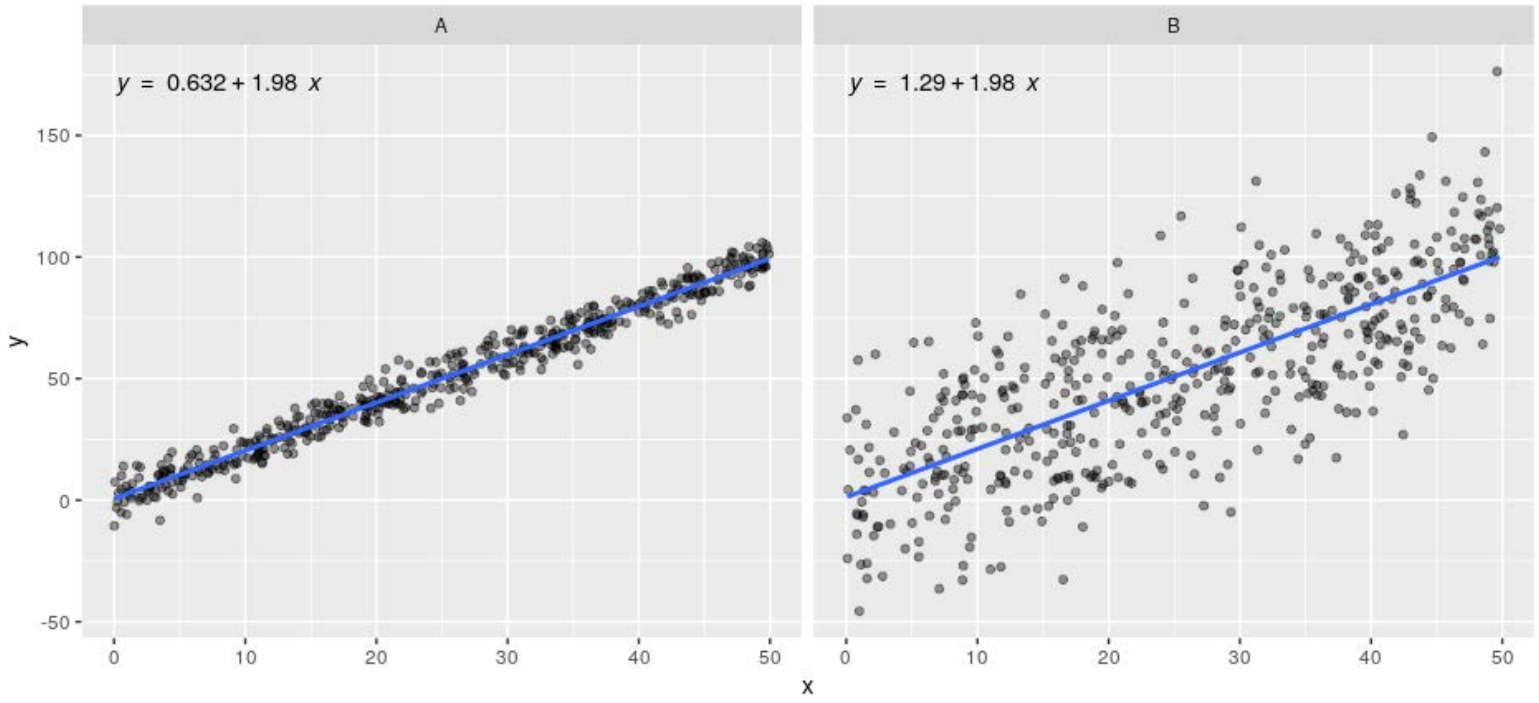
Test Yourself

Consider the regression of Y on X for two groups: A and B.

a. The slope of the regression relationship is larger for group A.

b. The slope of the regression relationship is larger for group B.

c. The slope of the regression relationship is the same between the two groups.



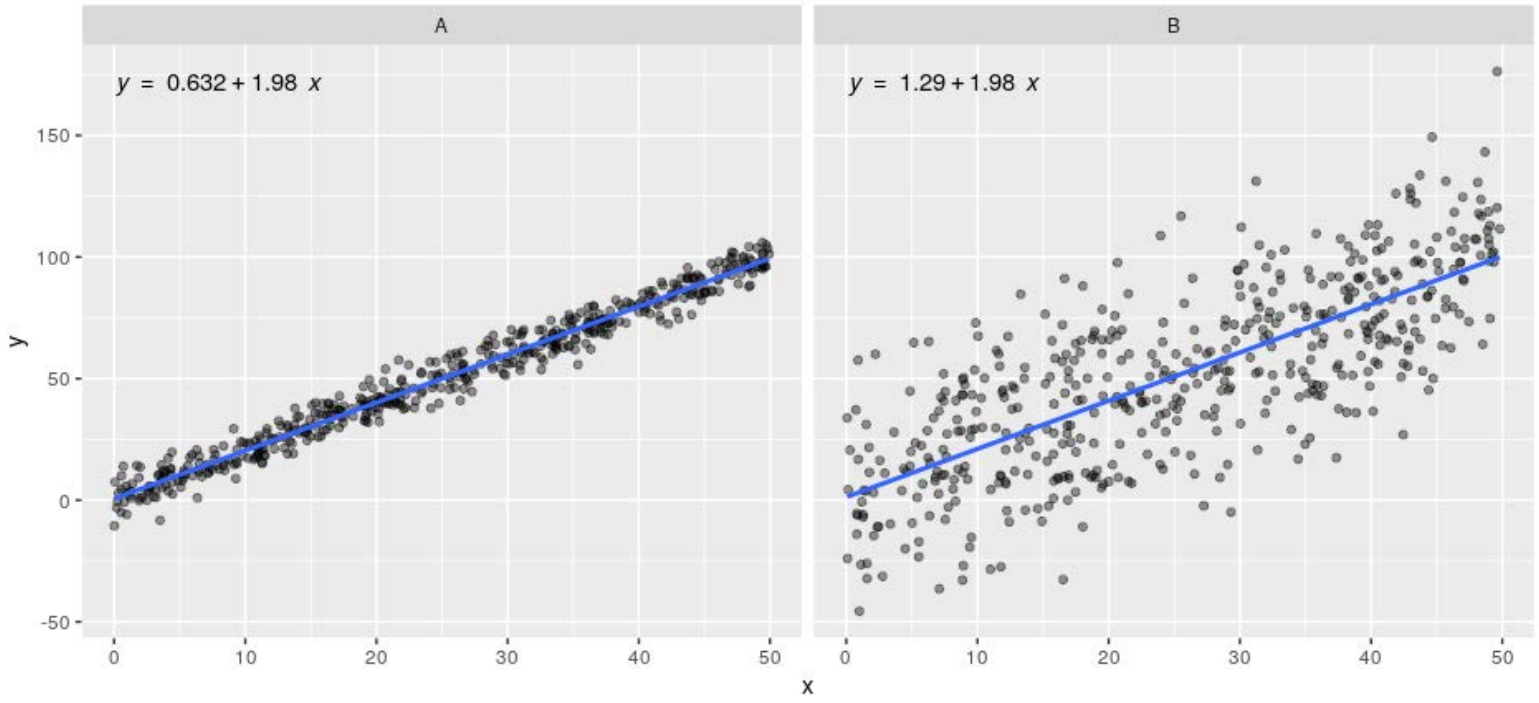
Test Yourself

Consider the regression of Y on X for two groups: A and B.

a. The magnitude of the correlation is larger for group A.

b. The magnitude of the correlation is larger for group B.

c. The magnitude of the correlation is the same between the two groups.



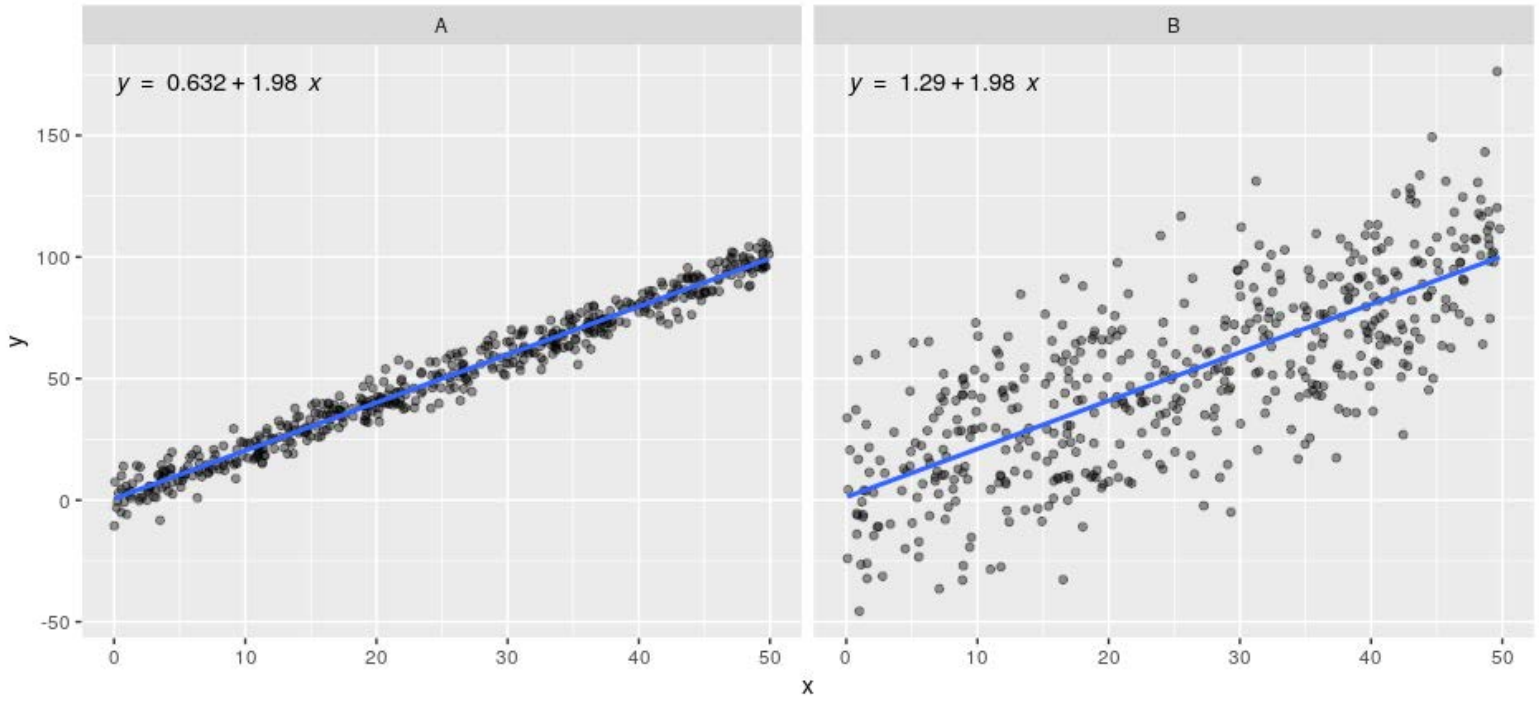
Test Yourself

Consider the regression of Y on X for two groups: A and B.

a. The magnitude of the correlation is larger for group A.

b. The magnitude of the correlation is larger for group B.

c. The magnitude of the correlation is the same between the two groups.



Some problems with regression/correlation

- Range restrictions – the relationship can change depending on the range of values sampled
- Heterogenous subsamples – the relationship can change depending on who you sampled
- Violation of assumptions
- Establishing causation

The GMAT problem

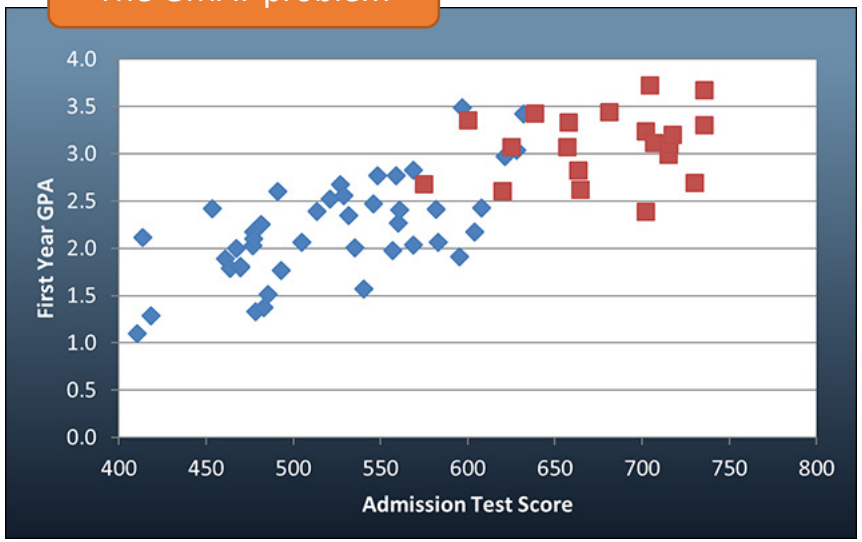
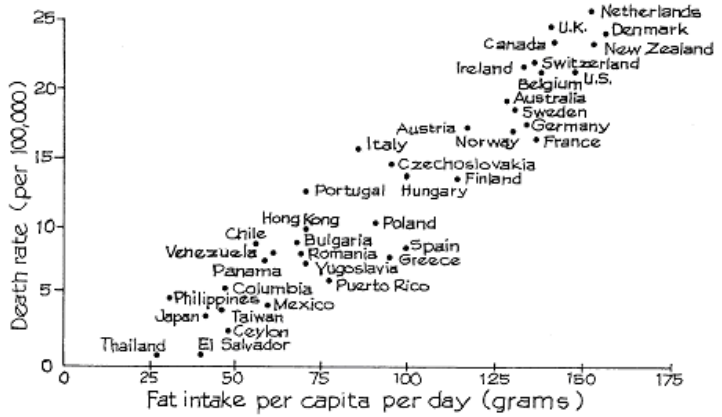


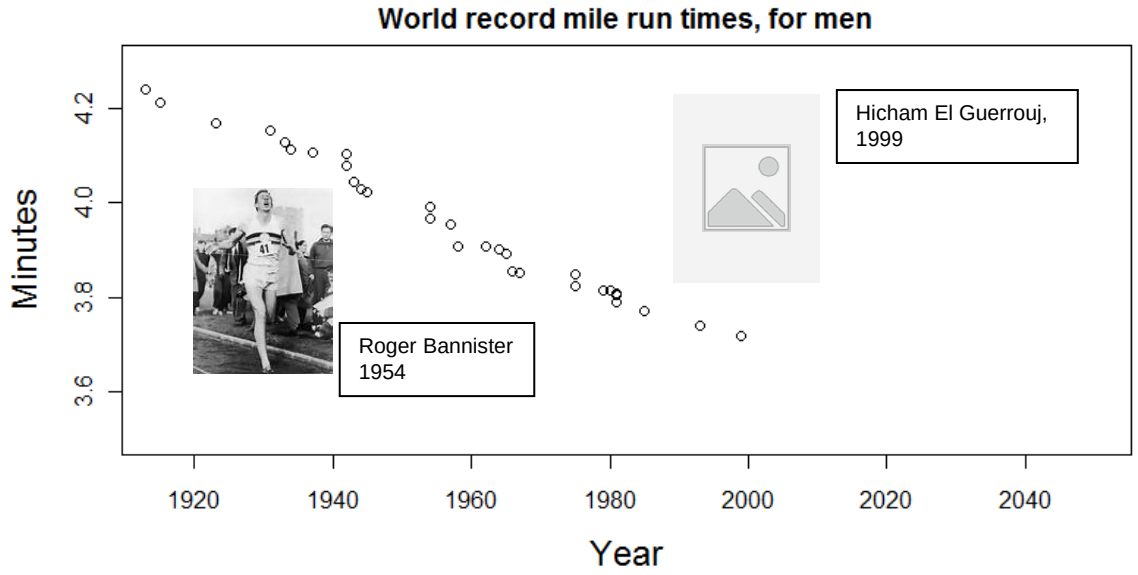
Figure 8. Cancer rates plotted against fat in the diet, for a sample of countries.



Source: K. Carroll, "Experimental evidence of dietary factors and hormone-dependent cancers," *Cancer Research* vol. 35 (1975) p. 3379. Copyright by *Cancer Research*. Reproduced by permission.

Some problems with regression/correlation

- Missing data may be a problem
  - Always make sure you subset to "complete cases"
  - There are more sophisticated approaches available (beyond the scope of this course)
  - Acknowledge the limitations present when there are missing data



Can you predict the world record time in 2040?

Answers to Questions

Q1a. $\hat{Y} = -17.60 + 3.93X$

Q1b. Yes. β_1-3.9, p<.001

Q1c. The predicted stopping distance is -17.6 feet.

Q2. (1.887082 + 50*0.276702)^2 = 247.2 mph

Packages and Functions

- `lm`
- `boxcox`
- `cor.test`
- `anova`
- `predict`
- `confint`
- `residuals`