IMS 24: Inference for linear regression with a single predictor

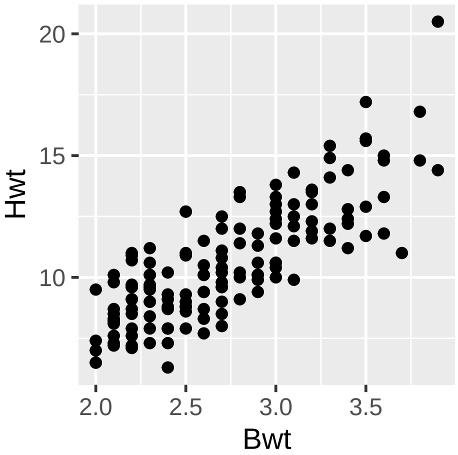
Packages

```
library(MASS)       # for data set
library(tidyverse) # for ggplot functions/plotting
library(broom)      # for tidy() function
```

Example

- Larger heart weights indicate a higher risk of heart attacks/disease in cats; however, heart weight is hard to measure.
- Want to see if there is a relationship between heart weight (`Hwt`) and body weight (`Bwt`) for domestic cats.
- If so, we will have a better idea of which cats are at risk for heart attacks/disease.

Example



Fitting a line to data

Recall:

$$y = \beta_0 + \beta_1 x + e$$

- β_0: intercept
- β_1: slope
- x: **predictor** variable
- y: **response** variable
- e: error (source of wiggliness around the line)

Inference for slope

Is our predictor (body weight) a good indicator for the response (heart weight)?

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

We learned how to get our estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ earlier in the course.

- $\hat{\beta}_1 = r\frac{s_y}{s_x}$
- $\hat{\beta}_0 = \overline{y} - b_1\overline{x}$

...or we can use R.

Example

```
fit <- lm(Hwt ~Bwt, data = cats)
summary(fit)
```

```
Call:
lm(formula = Hwt ~ Bwt, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5694 -0.9634 -0.0921  1.0426  5.1238 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -0.3567     0.6923  -0.515    0.607    
Bwt           4.0341     0.2503  16.119   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.452 on 142 degrees of freedom
Multiple R-squared:  0.6466,    Adjusted R-squared:  0.6441 
F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

Example

```r
fit <- lm(Hwt ~Bwt, data = cats)
summary(fit)
```

Data Set

Response variable

Predictor Variable

```
Call:
lm(formula = Hwt ~ Bwt, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5694 -0.9634 -0.0921  1.0426  5.1238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3567     0.6923  -0.515    0.607
Bwt           4.0341     0.2503  16.119   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.452 on 142 degrees of freedom
Multiple R-squared:  0.6466,    Adjusted R-squared:  0.6441
F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

$\hat{\beta}_0$ Intercept

$\hat{\beta}_1$ Slope

Interpretation

Interpreting Coefficients

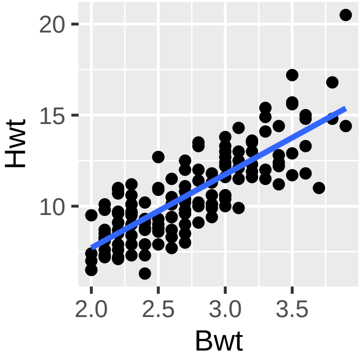
- The expected mean value for Y when $X = 0$ is $\hat{\beta}_0$
- For every one x-unit increase in X we expect the mean value of Y to change by $\hat{\beta}_1$ y-units

Interpretations should be in context of the problem. For example:

- The expected mean value for heart weight (grams) for cats when body weight (kg) is 0 is -0.3567.
- For every 1 kg increase in body weight we expect the mean value of heart weight to increase by 4.0341 grams

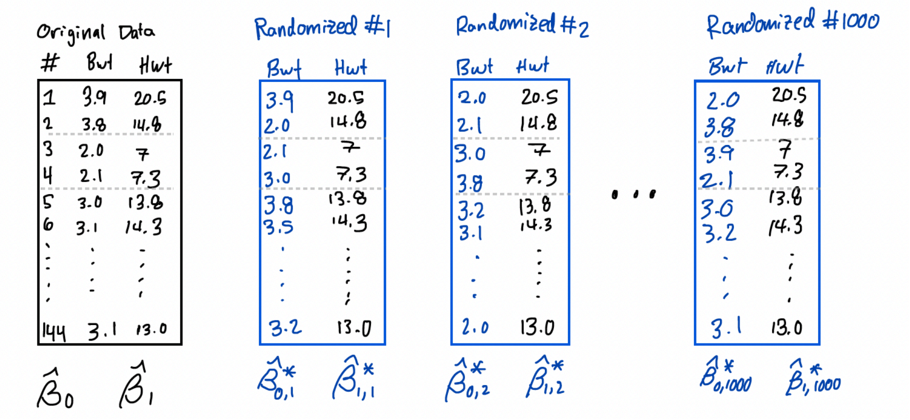
Example

```
ggplot(cats, aes(x = Bwt, y = Hwt))+
  geom_point()+
  geom_smooth(se = F, method = "lm")
```

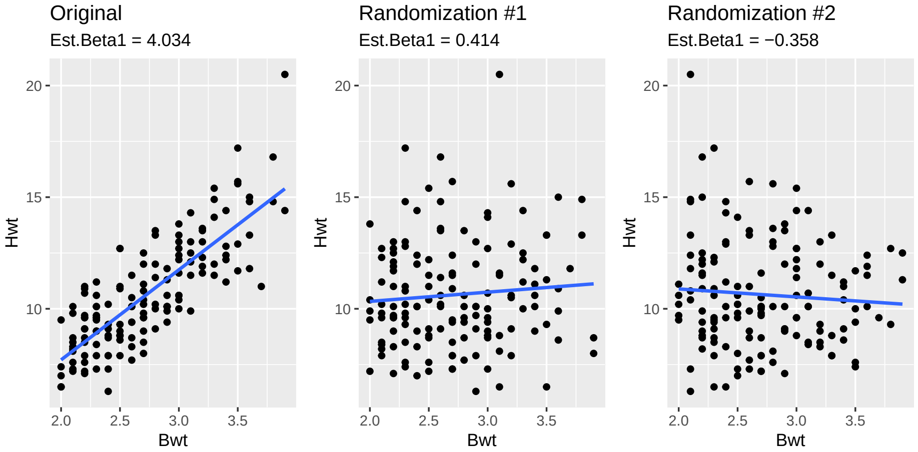


- The regression model merely approximates the true relationship between x and y
- The real relationship will not be exactly linear.
- If we had a slightly different data set our line would change.

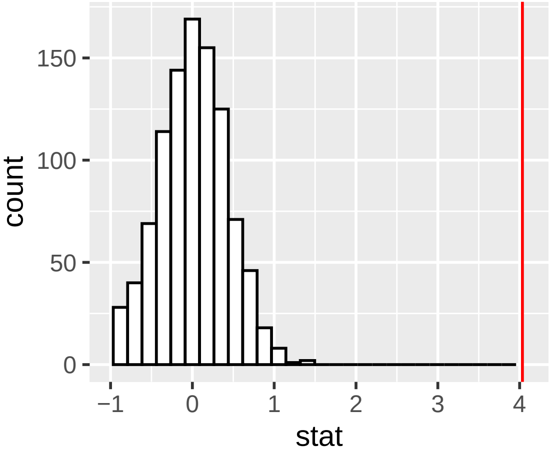
Randomization Test



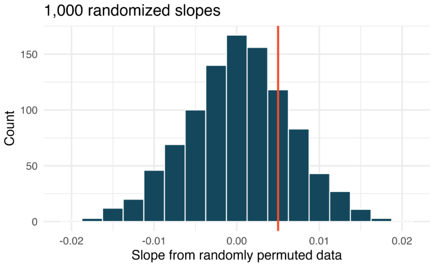
Randomization Test



Randomization Test



Practice Problem



[IMS 24.9] Baby's weight and father's age, randomization test. US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1000 births from 2014. Here, we study the relationship between the father's age and the weight of the baby.

a. What are the null and alternative hypotheses for evaluating whether the slope of the model for predicting baby's weight from father's age is different than 0?
b. The histogram describes the distribution of slopes when the null hypothesis is true. Use this histogram find the p-value and conclude the hypothesis test in the context of the problem.

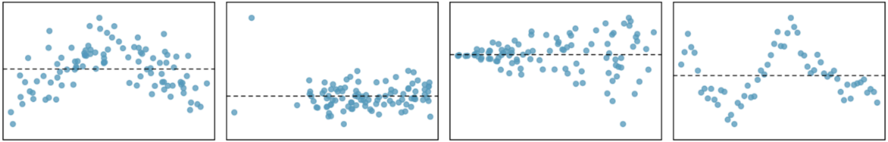
Assumptions for Mathematical Model

Assumptions need for inference for β_1

- **Linearity**. The data should show a linear trend.
- **Independent observations**. Be cautious about applying regression to data, which are sequential observations in time.
- **Nearly normal residuals**. Generally, the residuals must be nearly normal. Look for a random dismemberment of points around the zero line of a residual plot.
- **Constant or equal variability**. The points in the residual plot should not have a distinct/changing pattern.

Checking Assumptions

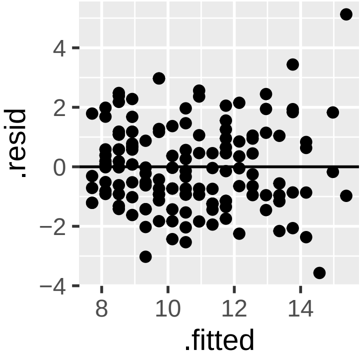
Residual Plots with problems



Example

Residual plot for cats data set

```
ggplot(fit, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0)
```



Test statistic for β_1

- Compute the standard error and the test statistic for $\hat{\beta}_1$.
- We can label the test statistic as T, because traditionally we rely on the t-distribution to test $\hat{\beta}_1$.

$$T = \frac{\hat{\beta}_1 - 0}{SE_{\hat{\beta}_1}}$$

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.3567	0.6923	-0.515	0.607	
Bwt	4.0341	0.2503	16.119	<2e-16	***
	$\hat{\beta}_1$	$SE_{\hat{\beta}_1}$	T	P-value	

Example

Is our predictor (body weight) a good indicator for the response (heart weight)?

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

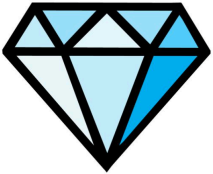
What is our conclusion?

The p-value for $\hat{\beta}_1$ is approximately 0 which is less than $\alpha = 0.05$. We conclude against (reject) H_0. The true value for β_1 is likely not 0.

Practice Problem

The `diamonds` data set in R contains the prices and other attributes of almost 54,000 diamonds. We want to see if `carat` (weight of the diamond) is a good predictor for `price` (in US dollars).

a) Write the hypotheses.
b) Check the conditions and comment on any potential violations.
c) What is the p-value and conclusion?



Constructing Confidence Intervals

We can construct confidence intervals for $\hat{\beta}_i$.

Most research focuses on $\hat{\beta}_1$.

Confidence Interval for $\hat{\beta}_i$

Let $i = 1$ or 0

- Lower $= \hat{\beta}_i - t^*_{df} SE_{\hat{\beta}_i}$
- Upper $= \hat{\beta}_i + t^*_{df} SE_{\hat{\beta}_i}$

Constructing Confidence Intervals

```
lm(Hwt ~Bwt, data = cats) |>
  tidy(conf.int = TRUE, conf.level=.95)
```

```
# A tibble: 2 x 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)   -0.357     0.692    -0.515 6.07e- 1    -1.73      1.01
2 Bwt            4.03      0.250    16.1   6.97e-34     3.54      4.53
```

Constructing Confidence Intervals

Response Variable → `Hwt`; Predictor Variable → `Bwt`; Data Set → `cats`

```r
lm(Hwt ~Bwt, data = cats) |>
  tidy(conf.int = TRUE, conf.level=.95)
```

← Confidence Interval Level

```
# A tibble: 2 x 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)   -0.357     0.692    -0.515 6.07e- 1    -1.73      1.01
2 Bwt            4.03      0.250    16.1   6.97e-34     3.54      4.53
```

Confidence Interval Bounds for $\hat{\beta}_1$

Practice Problem

The `diamonds` data set in R contains the prices and other attributes of almost 54,000 diamonds. We want to see if `carat` (weight of the diamond) is a good predictor for `price` (in US dollars).

d) Interpret the slope in context.

e) Calculate a 90% confidence interval for the slope of carat.

f) Do your results from the hypothesis test and the confidence interval agree? Explain.