# 1 Review

| Question | point estimate | parameter of interest | expected value of the sampling distribution | variance of the sampling distribution | standard error | confidence interval |
|---|---|---|---|---|---|---|
| single proportion | $\hat{p}$ | $p$ | $p$ | $\frac{p(1-p)}{n}$ | $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ | $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ |
| difference of two proportion | $\hat{p}_1 - \hat{p}_2$ | $p_1 - p_2$ | $p_1 - p_2$ | $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$ | $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ | $(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ |
| single mean (dependent samples, paired data) | $\bar{x}$ | $\mu$ | $\mu$ | $\frac{\sigma^2}{n}$ | $\frac{s}{\sqrt{n}}$ | $\bar{x} \pm t^*_{df} \frac{s}{\sqrt{n}}$ |
| difference of two means | $\bar{x}_1 - \bar{x}_2$ | $\mu_1 - \mu_2$ | $\mu_1 - \mu_2$ | $\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ | $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ | $(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ |

Hypothesis Testing

standard score (z or t) = $\frac{\text{point estimate} - \text{null value}}{StandardError}$

## 2 Simple Linear Regression

You may recall from your high school algebra class (and your calculus class) the equation of a line as

$y = mx + b$ where $m$ represents the slope of the line and $b$ represents the y-intercept.

In statistics we try to explain the relationship between two continuous variables using a linear regression model (if certain conditions are met).

The equation for a simple linear regression model is as follows:

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

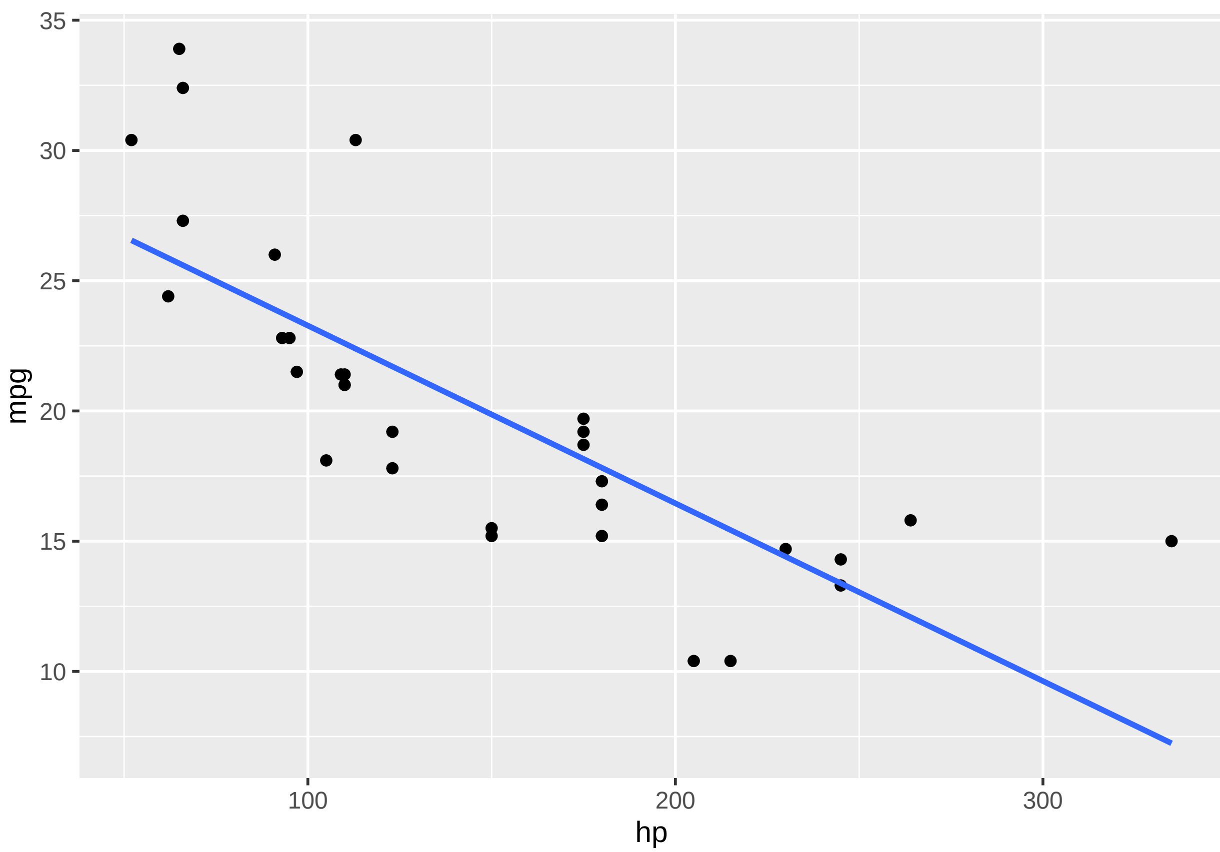
| description | point estimate | parameter of interest | Hypotheses |
|---|---|---|---|
| intercept | OpenIntro: $b_0$<br>Other resources: $\hat{\beta}_0$ | $\beta_0$ | $H_0 : \beta_0 = 0$ |
| slope | OpenIntro: $b_1$<br>Other resources: $\hat{\beta}_1$ | $\beta_1$ | $H_0 : \beta_1 = 0$ |

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

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Error/residual: $e = y - \hat{y}$

```r
mtcars %>%
  ggplot(aes(x = hp, y = mpg)) +
  geom_point() +
  geom_smooth(method = 'lm', se = FALSE)
```

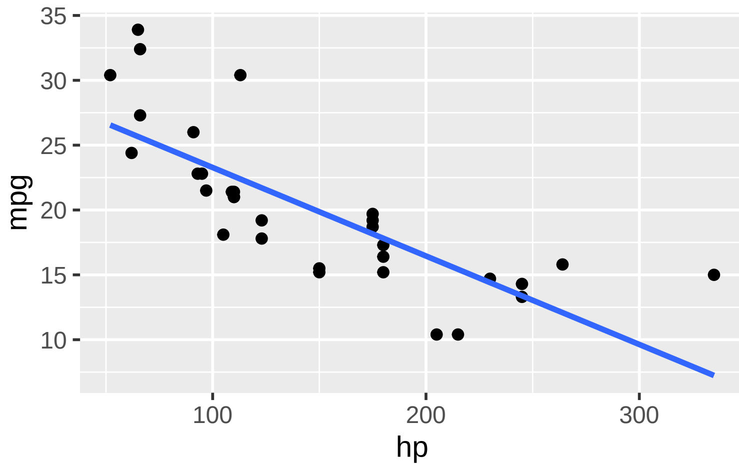


```r
lm(mpg ~ hp, data = mtcars) %>%
  summary()
```

```
##
## Call:
## lm(formula = mpg ~ hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7121 -2.1122 -0.8854  1.5819  8.2360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
## hp          -0.06823    0.01012  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.863 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

Understanding the R output

Residuals




```
mtcars %>%
  select(mpg, hp) %>%
  slice(1)
```

```
##   mpg  hp
## 1  21 110
```

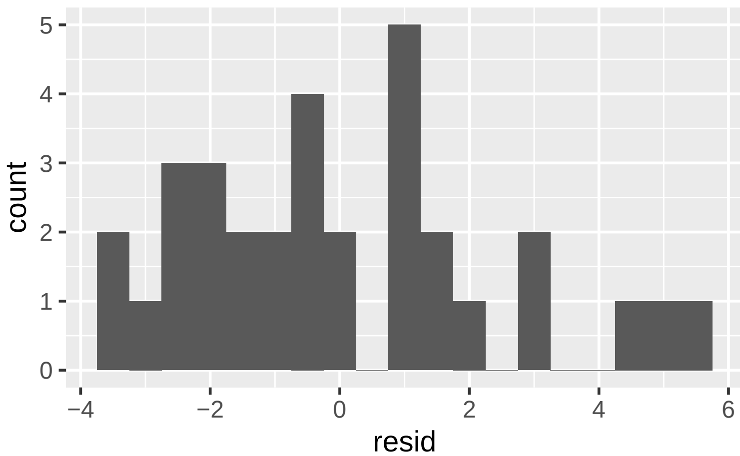
```
mtcars %>%
  select(mpg, hp) %>%
  slice(18)
```

```
##    mpg hp
## 1 32.4 66
```

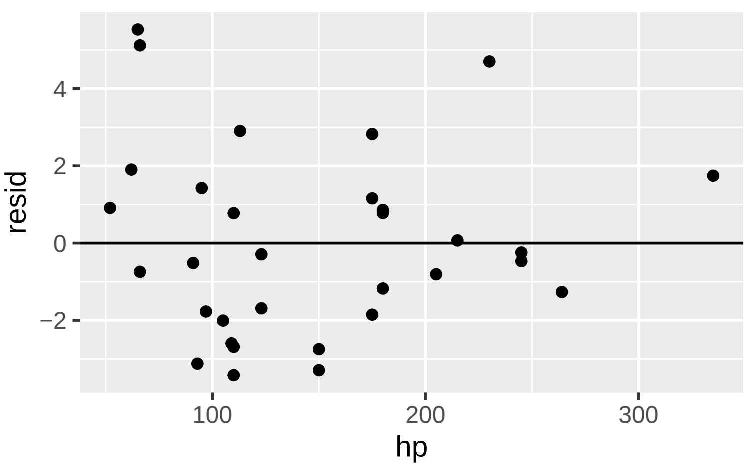
## 2.1 Estimation

Conditions

1. Linearity: The relationship between x and y has to be linear.
2. Independent Observations
3. Normality of Residuals




4. Constant Varibility



# 3 Multiple Linear Regression

Equation for multiple linear regression is

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 .... + \beta_k x_k + \epsilon$

where $k$ is the number of predictors.

```
mtcars %>%
  select(mpg, hp, am, wt) %>%
  glimpse()
```

```
## Observations: 32
## Variables: 4
## $ mpg <dbl> 21.0, 21.0, 22.8, 21.4, 18.7, 18.1, 14.3, 24.4, 22.8, 19.2...
## $ hp  <dbl> 110, 110, 93, 110, 175, 105, 245, 62, 95, 123, 123, 180, 1...
## $ am  <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1...
## $ wt  <dbl> 2.620, 2.875, 2.320, 3.215, 3.440, 3.460, 3.570, 3.190, 3....
```

```
lm(mpg ~ hp + am + wt, data = mtcars) %>%
  summary()
```

```
##
## Call:
## lm(formula = mpg ~ hp + am + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## am           2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```