RESIDUALS:

The mean of the residuals is 0
The mean of the fitted values is the same as the mean of original Y values
Yhat = b0 + b1*X
Cor(Yhat, X) = -1
Cor(e, X) = 0
Y = Yhat + e

Var(Y) = Var(Yhat) + Var(e)
SST = SSR + SSE
Total sum of squares = Regression ss + error ss
SST: How much variation there is in Y
SSR: How much of the variation is explained by X
SSE: How much of the variation is not explained by X
For a good fit, we hope for really large SSR and really small SSE.
SST is a fixed value that does not depend on X.
SSR and SSE do.

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.415  -3.429   0.212   3.333  11.908

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.36966    2.07326   -1.14     0.26
mydata$feet  0.05008    0.00303   16.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.03 on 34 degrees of freedom
Multiple R-squared:  0.889,     Adjusted R-squared:  0.886
F-statistic:  273 on 1 and 34 DF,  p-value: <2e-16

---
```

$\frac{b_1}{s_{b_1}}$

$s_e = \sqrt{\frac{SSE}{n-k-1}}$

$R^2 = \frac{SSR}{SST}$

COEFFICIENT OF DETERMINATION: R^2

$R^2 = SSR/SST = 1 - (SSE/SST)$
R^2 is between 0 and 1, and the closer R^2 is to 1, better the fit.
1 does not mean predictions are good. It means you have modeled the variation well.
You can have a low R^2 value for a good model and high R^2 value for a bad model.

REGRESSION:

Useful for linear relationships.
We model the average of something rather than the something itself.
The regression line should be viewed as the average value of Y for a given X, or in symbols E(Y|X).
$E(Y|X) = \beta_0 + \beta_1 X$ or $Y = \beta_0 + \beta_1 X + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$
$\beta_0 + \beta_1 X$: The part of Y related to X
ε: The part of Y unrelated to X i.e. regression error

ESTIMATING VARIANCE AND INTERVALS:

$S_e^2 = SSE/(n-k-1)$: our estimate of σ^2

n: number of rows in dataset

k: number of variables in model

95% of the Y values should lie within the interval
$\hat{Y} \pm 1.96 s_e$ or $b_0 + b_1 X \pm 1.96 s_e$
95% of the β_0 values should lie within the interval
$b_0 \pm 1.96(Sb_0)$
95% of the β_1 values should lie within the interval
$b_1 \pm 1.96(Sb_1)$

β_0 and β_1 don't mean much if the confidence intervals spans across 0.

```
Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) 17066.76607   169.02464   101.0   <2e-16 ***
Odometer       -0.06232     0.00462   -13.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 303 on 98 degrees of freedom
  (139 observations deleted due to missingness)
Multiple R-squared:  0.65,      Adjusted R-squared:  0.647
F-statistic:  182 on 1 and 98 DF,  p-value: <2e-16
```

We are roughly 95% confident that the (average) price of an Accord with 50,000 miles is in the interval

$$17066 - 0.06(50000) \pm 1.96(303) = (13472, 14660)$$

s_{b_0}

```
Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) 17066.76607   169.02464   101.0   <2e-16 ***
Odometer       -0.06232     0.00462   -13.5   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

s_{b_1}

HYPOTHESIS TESTS FOR REGRESSION MODEL:

$$T = \frac{b_1 - \beta_1^*}{s_{b_1}}$$

$H_0 : \beta_1 = \beta_1^*$ If $|T| > 1.96$ *reject* H_o

$H_a : \beta_1 \neq \beta_1^*$

$H_0 : \beta_1 = \beta_1^*$ If $T < -1.64$ *reject* H_o

$H_a : \beta_1 < \beta_1^*$

$H_0 : \beta_1 = \beta_1^*$ If $T > 1.64$ *reject* H_o

$H_a : \beta_1 > \beta_1^*$

There are then three ways one could test this hypothesis; get familiar with at least one:

- ❑ Test statistic — $|t| > 1.96$ — reject
- ❑ P-value — p-val < 0.05. — reject
- ❑ Confidence Interval — if 0 not in conf. interval reject

To test whether *X* affects Y, test whether $\beta_1 = 0$

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

The test statistic is

$$t = \frac{b_1 - 0}{s_{b_1}} = \frac{114 - 0}{15.2} = 7.5$$

	Estimate	Std. Error	t value
(Intercept)	845.2	48.3	17.52
mydata$GPA	114.0	15.2	7.52

and

$7.5 > 1.96$

so we reject the null hypothesis.

```
Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept)  11253.1     1170.8    9.61 0.000000000000015 ***
mydata$mpg    -238.9       53.1   -4.50 0.000025461312051 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2620 on 72 degrees of freedom
Multiple R-squared:  0.22,      Adjusted R-squared:  0.209
F-statistic: 20.3 on 1 and 72 DF,  p-value: 0.0000255
```

Interpretation? Ho : β1 = 0
H1: β1 ≠ 0
p-value = 0.00002546 < 0.05 reject
| t | = 4.5 > 1.96 reject

Reject: variable is needed in the model

STANDARDIZED RESIDUALS:

$$r_i = \frac{e_i}{s_e} \approx \frac{\varepsilon_i}{\sigma} \sim N(0,1)$$

Regression error ε by itself is dependent on units and that could be a problem.
Sometimes we use standardized residuals for convenience.

Plotting residuals vs yhat or standardized residuals vs yhat give us a blob(the two should not be related) for a good model. The plots look the same, just one standardized between ± 2
eg: Residual value of 10 is not an outlier but standardized residual value is.

OUTLIERS:

Unusual points in the x-space - leverage - these just move the line in a direction
Unusual points in the y-space - influential - these alter the slope of the line
Cook's distance: Calculated for each row of dataset. Extreme values indicate influential

TRANSFORMATIONS:

For linear regression, if Y vs X is not linear, you want to try to make it linear
Usual transformations: log(X), √X, X^2, 1/X. We typically transform X before Y to leave it interpretable

NORMALITY TEST:

Ho: Normal Ha: Not Normal
We want a high p-value i.e. we want our data to be normal. If not normal, a log transformation works most times.

ADJUSTED R^2:

R^2 goes up for every X added to the model.
R_a^2 adds a penalty and compensates for that.

$$R_a^2 = 1 - \frac{\frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2}{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\frac{1}{n-k-1}SSE}{\frac{1}{n-1}SST}$$

DUMMY VARIABLES:

Takes values of 0 or 1. Need n-1 variables for n categories. The one left out is 'baseline' and included in the intercept.

INTERACTION VARIABLES:

Formulated as the product of two variables e.g.: X3 = X1*X2

With two *x* variables the model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e$$

If we factor out x_1 we get:

$$y = \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 + e$$

If this variable in the model, it means level of X2 changes how X1 affects Y

HETEROSKEDASCTICITY:

Homoskedastic noise: $\varepsilon_i \sim N(0 , \sigma^2)$ → One term to estimate i.e. σ^2 → Constant variation

Heteroskedastic noise: $\varepsilon_i \sim N(0 , \sigma_i^2)$ → n terms to estimate i.e. σ_i^2 → Non-constant variation

We would rather have one term to estimate.

If you have heteroskedasaticity, your estimates are ok, but your standard errors are incorrect. Hypothesis testing will be wrong, confidence interval will be wrong.

An easy way to reduce the variation in Y is to take the log of it. → Can't compare Se to previous model after this because different units.

```
> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 5.938511    Df = 1     p =
0.01481353
```

Ho : It's homoskedastic

Ha: It's heteroskedastic

p is low, Ho must go

there is non constant variation in the noise, we must fix it

MULTICOLLINEARITY:

it is hoped that the explanatory variables will be highly correlated with the dependent variable.

At the same time, however, it is not desirable for strong relationships to exist among the explanatory variables.

When explanatory variables are correlated with one another, the problem of multicollinearity is said to exist.

Problem: The standard deviations of the regression coefficients (s_bi) will be disproportionately large. As a result, the t-ratios will be small. Thus we may think we do not need variables when in fact we do

Detecting: Compare the pairwise correlations between the explanatory variables. One rule of thumb is that multicollinearity may be a serious problem if any pairwise correlation is larger than 0.5

Solution: Drop the one with lesser significance to Y. Get more data. Redefine variables. Step-wise regression.