# Chapter 3: Linear Regression

*Linear regression* is a simple approach for supervised learning when the response is quantitative. Linear regression has a long history and we could actually spend most of this semester talking about it.

Although linear regression is not the newest, shiniest thing out there, it is still a highly used technique out in the real world. It is also useful for talking about more modern techniques that are **generalizations** of it. Ridge regression, Lasso, logistic regression, GAMs...

We will review some key ideas underlying linear regression and discuss the least squares approach that is most commonly used to fit this model.

Linear regression can help us to answer the following questions about our `Advertising` data:

1. Is there a relationship between advertising and sales?
   i.e. should people spend money on ads?
2. How strong is the relationship between ads & sales?
   i.e. how well can we predict sales based on ads?
3. Which media contribute to sales?
4. How accurately can we predict the effect of each medium on sales?
5. How accurately can we predict future sales?
6. Is the relationship linear?
7. Is there synergy among the advertising media?
   i.e. is $50k for TV and $50k for radio better than $100k on radio or TV alone?

# 1 Simple Linear Regression

*Simple Linear Regression* is an approach for predicting a quantitative response $Y$ on the basis of a single predictor variable $X$.

It assumes:

- approximately linear relationship between X and Y
- random error term is Normally distributed
- random error term has constant variance.

Which leads to the following model:

$$Y = \overbrace{\beta_0 + \beta_1 X}^{\text{linear relationship}} + \varepsilon$$

$$\varepsilon \sim \underbrace{N(0, \sigma^2)}_{\text{error assumptions}}$$

For example, we may be interested in regressing `sales` onto `TV` by fitting the model

$$\text{sales} = \underline{\beta_0} + \underline{\beta_1} * TV + \varepsilon$$

unknown constants (intercept & slope)
"parameters", "model coefficients"

^ "hat" = estimates or predictors

Once we have used training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we can predict future sales on the basis of a particular TV advertising budget.

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

particular value of $X = x$.

prediction of $y$

## 1.1 Estimating the Coefficients

In practice, $\beta_0$ and $\beta_1$ are **unknown**, so before we can predict $\hat{y}$, we must use our training data to estimate them.

"fit the model"

Let $(x_1, y_1), \dots, (x_n, y_n)$ represent $n$ observation pairs, each of which consists of a measurement of $X$ and $Y$.

In the advertising data,

$X$ = TV ad budget

$Y$ = sales

$(x_1, y_1), \dots, (x_{200}, y_{200})$ = training data from $n = 200$ markets.

**Goal:** Obtain coefficient estimates ~~beta0~~ $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the linear model fits the available data "well."

i.e. $y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i$ for $x_i = 1, \dots, n$

We want to find an intercept $\hat{\beta}_0$ and slope $\hat{\beta}_1$ s.t. the resulting line is "close" to the $n = 200$ points.

The most common approach involves minimizing the *least squares* criterion.

↳ there are other methods we will talk about in Ch. 6.

Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ prediction for $Y$ based on $i$th value of $X$.

$e_i = y_i - \hat{y}_i$ $i$th residual

$RSS = e_1^2 + \dots + e_n^2$ residual sum of squares.

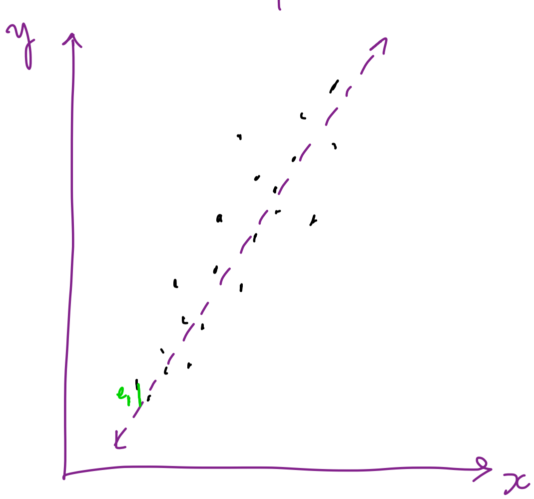
choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize RSS

How? $RSS = \sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2$

a function of $\hat{\beta}_0$ and $\hat{\beta}$ Want

$\underset{\hat{\beta}_0, \hat{\beta}_1}{\arg\min} \; RSS(\hat{\beta}_0, \hat{\beta}_1) \Rightarrow$ Calculus!

1. take derivatives
2. set $= 0$
3. solve for $\hat{\beta}_0, \hat{\beta}_1$




results in

The least squares approach results in the following estimates:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

"least squares coefficients"

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

We can get these estimates using the following commands in `R`:

```r
## load the data in
ads <- read_csv("../data/Advertising.csv")

## fit the model
model <- lm(sales ~ TV, data = ads)
```

formula for model $Y \sim X$ "regress $Y$ on $X$"

specify data frame.

```r
summary(model)
```

get results summary

$\hat{\beta}_0$

$\hat{\beta}_1$

```
##
## Call:
## lm(formula = sales ~ TV, data = ads)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

## 1.2 Assessing Accuracy

Recall we assume the *true* relationship between $X$ and $Y$ takes the form

f unknown, fixed

$$Y = f(X) + \varepsilon$$

$\varepsilon$ mean-zero random term.

If $f$ is to be approximated by a linear function, we can write this relationship as

average increase in $Y$ associated w/ 1 unit increase in $X$

population regression line $\qquad Y = \beta_0 + \beta_1 X + \varepsilon$ catch-all term for what we miss w/ this simple model – true relationship may not be linear, may be other variables that explain variability in $Y$, measurement error, etc.

expected value of $Y$ when $X=0$

and when we fit the model to the training data, we get the following estimate of the *population model*

least squares line. $\qquad Y = \hat{\beta}_0 + \hat{\beta}_1 X$

But how close this this to the truth? measure w/ standard error

$$Var(\hat{\beta}_0) = [SE(\hat{\beta}_0)]^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

$$Var(\hat{\beta}_1) = [SE(\hat{\beta}_1)]^2 = \sigma^2\left[\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

In general, $\sigma^2$ (var($\varepsilon$)) is not known, so we estimate it with the *residual standard error*, $RSE = \sqrt{RSS/(n-2)}$.

residual sum of squares.

We can use these standard errors to compute confidence intervals and perform hypothesis tests.

95% CI for $\beta_1$ : $\hat{\beta}_1 \pm 2SE(\hat{\beta}_1)$

95% CI for $\beta_0$ : $\hat{\beta}_0 \pm 2SE(\hat{\beta}_0)$

hypothesis test:

$H_0$: There is no relationship between $X$ and $Y$ $\iff$ $H_0: \beta_1 = 0$

$H_a$: There is a relationship between $X$ and $Y$ $\qquad H_a: \beta_1 \neq 0$

?: Is $\hat{\beta}_1$ far enough away from 0 to be confident it is non zero? How far is far enough? depends on $SE(\hat{\beta}_1)$.

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \sim t_{n-2} \Rightarrow$$ compute P(observing any number equal to $|t|$ or larger in abs value | $H_0$ true) = p-value. If small enough $\Rightarrow$ unlikely $H_0$ is true

$\Rightarrow$ reject.

Once we have decided that there is a significant linear relationship between $X$ and $Y$ that is captured by our model, it is natural to ask

> To what extent does the model fit the data?

The quality of the fit is usually measured by the *residual standard error* and the $R^2$ statistic.

**RSE**: Roughly speaking, the RSE is the average amount that the response will deviate from the true regression line. This is considered a measure of the *lack of fit* of the model to the data.

$R^2$: The RSE provides an absolute measure of lack of fit, but is measured in the units of $Y$. So, we don't know what a "good" RSE value is! $R^2$ gives the proportion of variation in $Y$ explained by the model. i.e. will be between 0 and 1.

Advertising data example

```
summary(model)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = ads)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

$Y=$ $X=$

$\hat\beta_0$ $\hat\beta_1$ $SE(\hat\beta_0)$ $SE(\hat\beta_1)$

$H_0: \beta_i = 0$ $H_a: \beta_i \neq 0$ $i=0,1.$

p-value

RSE

$R^2$ = proportion of variability in $Y$ explained by linear relationship w/ $X$.

# 2 Multiple Linear Regression

Simple linear regression is useful for predicting a response based on one predictor variable, but we often have **more than one** predictor.

> How can we extend our approach to accommodate additional predictors?

We could run separate SLR for each predictor.

But how to make a single prediction for y based on levels of all predictors?

Also each model would ignore the other predictors... what if they are related?

↳ misleading results.

Solution! We can give each predictor a separate slope coefficient in a single model.

association with response

predictor

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

We interpret $\beta_j$ as the "average effect on $Y$ of a one unit increase in $X_j$, *holding all other predictors fixed*".

In our `Advertising` example,

$$\text{sales} = \beta_0 + \beta_1 \, \text{TV} + \beta_2 \, \text{radio} + \beta_3 \, \text{newspaper} + \varepsilon.$$

## 2.1 Estimating the Coefficients

As with the case of simple linear regression, the coefficients $\beta_0, \beta_1, \ldots, \beta_p$ are unknown and must be estimated. Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p.$$

now instead of a line, we are fitting a hyperplane.

The parameters are again estimated using the same least squares approach that we saw in the context of simple linear regression.

```
# model_2 <- lm(sales ~ TV + radio + newspaper, data = ads)
model_2 <- lm(sales ~ ., data = ads[, -1])

summary(model_2)
```

2 ways to fit same model.

"every other column in data"

```
##
## Call:
## lm(formula = sales ~ ., data = ads[, -1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

## 2.2 Some Important Questions

When we perform multiple linear regression we are usually interested in answering a few important questions:

1. Is atleast one of the predictors $X_1, \dots, X_p$ useful in predicting the response?

2. Do all the predictors help to explain $Y$, or is only a subset useful?

3. How well does the model fit the data?

4. Given a set of predictor values what response should we predict and how accurate is that prediction?

### 2.2.1 Is there a linear relationship between response and predictors?

We need to ask whether all of the regression slope coefficients are zero, which leads to the following hypothesis test.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a : \text{at least one } \beta_j \text{ is nonzero.}$$

This hypothesis test is performed by computing the $F$-statistic

variance explained by the model

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F_{p, n-p-1}$$

(under the null hypothesis $H_0$).

variance unexplained

If this ratio is large (much larger than 1), evidence against the null $H_0$, evidence there is some relationship.

### 2.2.2 Deciding on Important Variables

After we have computed the $F$-statistic and concluded that there is a relationship between predictor and response, it is natural to wonder

> Which predictors are related to the response?

We could look at the $p$-values on the individual coefficients, but if we have many variables this can lead to false discoveries.

Instead we could consider *variable selection*. We will revisit this in Ch. 6.

→ forward selection, backwards selection, Lasso

### 2.2.3 Model Fit

Two of the most common measures of model fit are the RSE and $R^2$. These quantities are computed and interpreted in the same way as for simple linear regression.

Be careful with using these alone, because $R^2$ will **always increase** as more variables are added to the model, even if it's just a small increase.

could lead to overfitting.
how to avoid? Use test data! Ch. 5.

```r
# model with TV, radio, and newspaper
summary(model_2)
```

individual p-values.

```
##
## Call:
## lm(formula = sales ~ ., data = ads[, -1])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## TV           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

$R^2$

F-test

$H_0: \beta_1 = \cdots = \beta_p = 0$

$H_a: \beta_j \neq 0 \quad j \in 1, \ldots, p.$

```r
# model without newspaper
summary(lm(sales ~ TV + radio, data = ads))
```

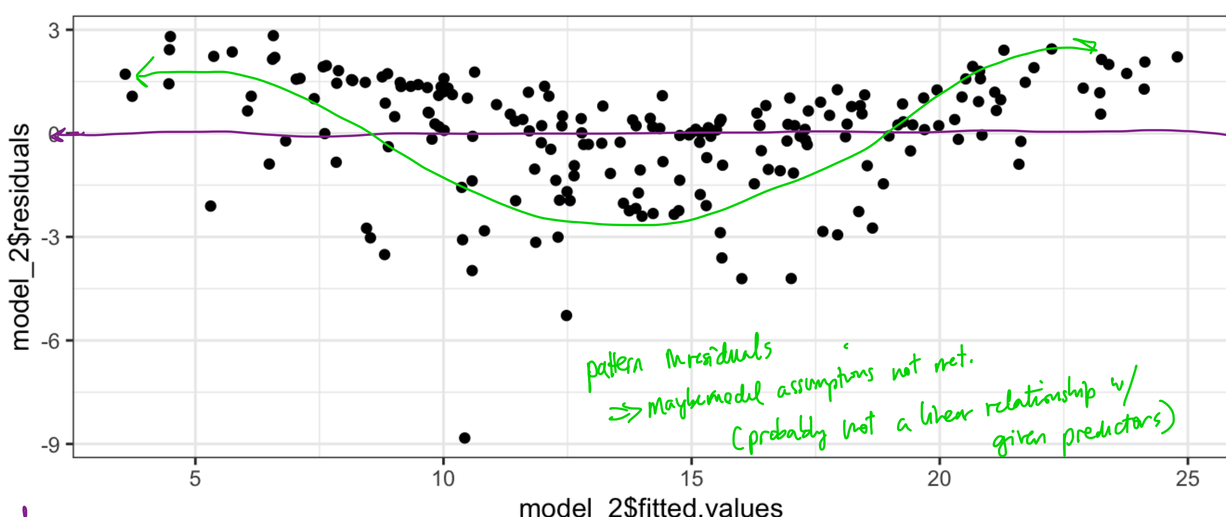
```
##
## Call:
## lm(formula = sales ~ TV + radio, data = ads)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7977 -0.8752  0.2422  1.1708  2.8328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92110    0.29449   9.919   <2e-16 ***
## TV           0.04575    0.00139  32.909   <2e-16 ***
## radio        0.18799    0.00804  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.681 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

$R^2$ barely decreased $\Rightarrow$ newspaper not contributing much to understanding variability in sales.

It may also be useful to plot residuals to get a sense of the model fit. (are our assumptions met?).

$e_i = y_i - \hat{y}_i$

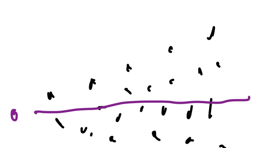
```r
ggplot() +
  geom_point(aes(model_2$fitted.values, model_2$residuals))
```

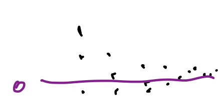



Want: random noise centered at zero, no pattern.

$\hat{y}$

Could also compare residuals to Normal dsn via QQ plots.




non constant variance $\sigma^2$ in errors.