# Statistics 452: Statistical Learning and Prediction

## Chapter 3, Part 1: Simple Linear Regression

Brad McNeney

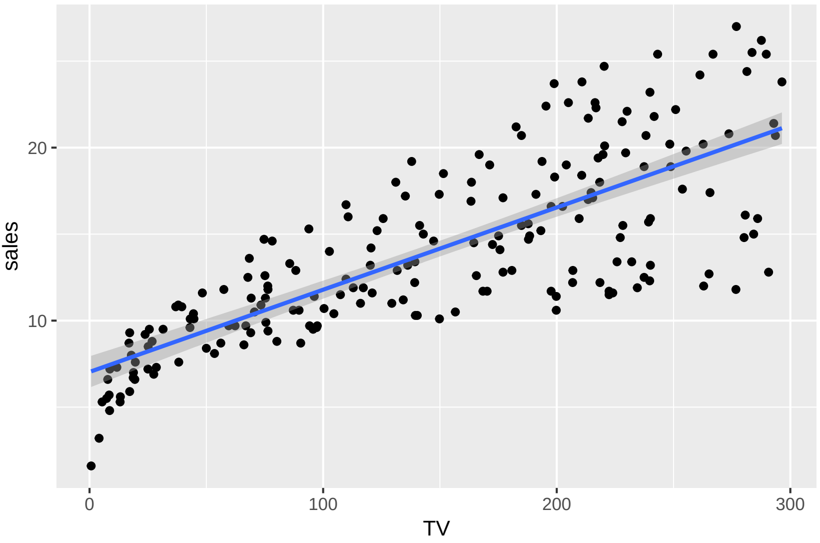
# Example: Advertising Data

- Sales (in thousands of units), and advertising budgets in thousands of dollars for TV, radio and newspaper for 200 markets.

```r
uu <- url("http://faculty.marshall.usc.edu/gareth-james/ISL/Advertising.csv")
advert <- read.csv(uu,row.names=1)
head(advert)
```

```
##      TV radio newspaper sales
## 1 230.1  37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 151.5  41.3      58.5  18.5
## 5 180.8  10.8      58.4  12.9
## 6   8.7  48.9      75.0   7.2
```

```r
ggplot(advert,aes(x=TV,y=sales)) + geom_point() +
  geom_smooth(method="lm")
```



# Simple Linear Regression Model

- Recall our general model from Chapter 2:

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

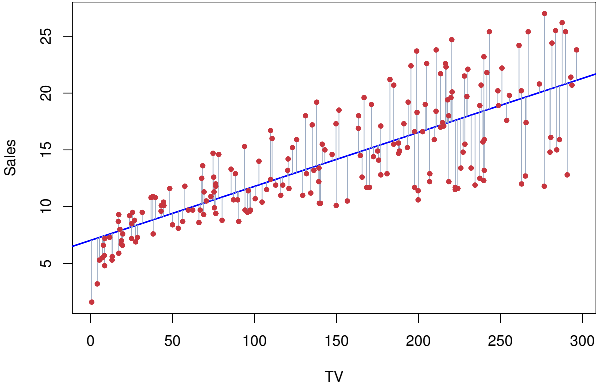
- Simple linear regression assumes the function $f$ is linear in a single predictor $X$; i.e., $f(X) = \beta_0 + \beta_1 X$.
  - $\beta_0$ is the intercept and
  - $\beta_1$ is the slope.

# Fitting the line

- We use the method of least squares to fit the line.
- Goal: Using observed data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ fit the model
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
where $\hat{y}_i$ is the *predicted* or *fitted value* of $Y$ for $X = x_i$.
- Idea: try all possible $\hat{\beta}_0$ and $\hat{\beta}_1$ until you find the line that fits the data the "best"; i.e. the $\hat{y}$'s are as close to the $y$'s as possible.
  - What is the criteria for best?

# Residuals

- The vertical distances $e_i = y_i - \hat{y}_i$ are called the residuals



# Least Squares

- Least squares minimizes the sum of the squared residuals, known as the **residual sum of squares**

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2$$

- There are many visual demonstrations of the least squares idea on the internet; e.g., http://www.dangoldstein.com/regression.html
  - Try clicking the $-$ slope, $+$ slope, $-$ intercept, and $+$ intercept buttons to minimize the sum of squared distances, summarized by the blue square.
  - Then click “Fit and lock” to see the line that minimizes the sum of squares.

# Least-Squares Regression

- The line that minimizes RSS has

$$\begin{aligned}\hat{\beta}_1 &= r\frac{s_y}{s_x} \\ \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1\overline{x}\end{aligned}$$

  where
  - $r$ is the Pearson correlation between the $x$'s and $y$'s,
  - $s_y$ is the sample SD of the $y$'s and $s_x$ is the sample SD of the $x$'s.
- We will always use R to calculate these estimates.

# Advertising Example

```r
afit <- lm(sales ~ TV,data=advert)
summary(afit)$coefficients
```

```
##               Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 7.03259355 0.457842940 15.36028 1.40630e-35
## TV          0.04753664 0.002690607 17.66763 1.46739e-42
```

# Accuracy of the Coefficient Estimates

- Least squares is a good way to fit a line to a scatterplot, but if we want to assess the accuracy of the coefficient estimates we need assumptions about the distribution of errors.
- Recall the errors $\epsilon$ in $Y = f(X) + \epsilon$.
- Errors are assumed to be normally distributed with mean zero and SD $\sigma$.
  - The $\epsilon$ are the irreducible error terms, and $\sigma$ quantifies the irreducible error.
- The SD of the error terms is assumed to be constant for all $x$.

# Model Summary

- We can summarize the model assumptions by saying that:
  - the $(X, Y)$ pairs are independent,
  - conditional on $X = x$, $Y$ has a normal distribution $N(f(x), \sigma)$, with conditional mean $f(x) = \beta_0 + \beta_1 x$ and conditional standard deviation $\sigma$ being a constant value (i.e. same for all $x$).

# SD and SE of Coefficient Estimators

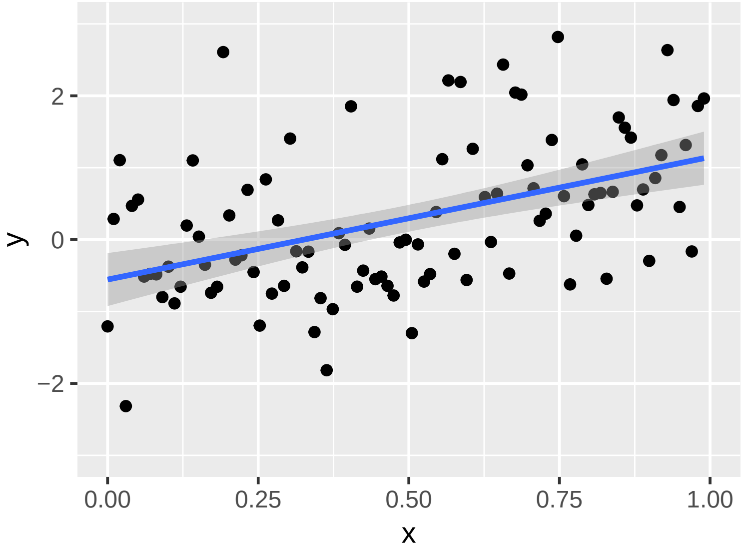
- Under the model assumptions, one can derive expressions for the SD of the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$, which the text refers to as the standard error (SE).
  - Side note: What the text calls the SE is what I call the SD, and what the text calls the estimated SE is what I call the SE. I'll try to stick to their terminology, but may slip.
- One can derive expressions for the SE and estimated SE.
  - E.G., equation (3.8) of text, page 66.
  - Both SE and estimated SE denoted $SE(\hat{\beta}_i)$.
  - We will always use the computer to estimate SEs.

# Simulation Example

- Start R on your computer, choose your own random seed and run the following sequence of code chunks.

```
# simulation parameters
n <- 100; beta0 <- 0; beta1 <- 1; sd <- 1; NREPS <- 1000
x <- seq(from=0,to=1,length=n)
#simulation function
simdat <- function() { # R finds sim params from workspace
  f <- beta0 + beta1*x
  y <- f + rnorm(n,mean=0,sd=sd)
  return(list(dat = data.frame(x=x,y=y),coef=coefficients(lm(y~x))))
}
# Set a random seed for replicability
set.seed(1234)
```

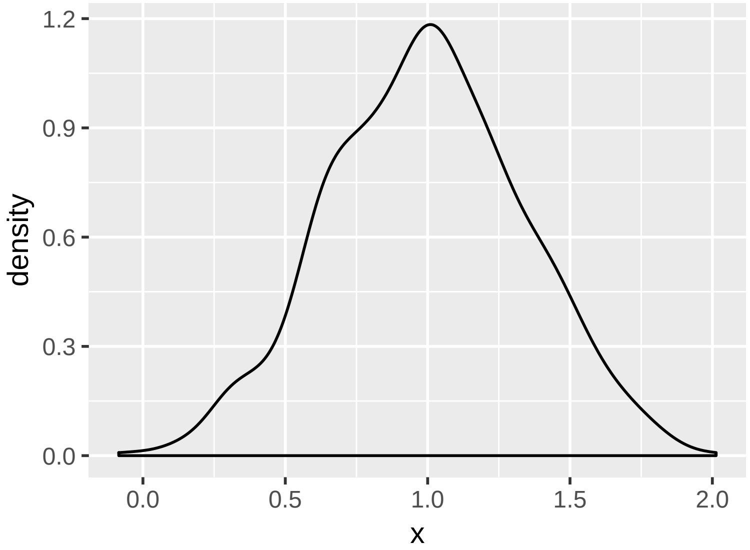
```r
# Do the following a few times
dat <- simdat()$dat
ggplot(dat,aes(x=x,y=y)) + geom_point() + geom_smooth(method="lm") + ylim(-3,3)
```



```r
simcoef <- function() {
  return(simdat()$coef)
}
simout <- replicate(NREPS,simcoef())
simout <- data.frame(t(simout))
head(simout)
```

```
##   X.Intercept.         x
## 1  -0.03538824 1.1532628
## 2   0.31615149 0.6769044
## 3   0.04367708 0.8964356
## 4  -0.32101657 1.5984614
## 5   0.01969292 0.6868602
## 6   0.01532160 0.7936332
```

```r
ggplot(simout,aes(x=x)) + geom_density()
```



# Confidence Intervals

- The sampling distribution of the coefficients can be used to derive the following probability statement:
  - There is a 95% chance that the interval

    $$(\hat{\beta}_1 - t^* SE(\hat{\beta}_1), \hat{\beta}_1 + t^* SE(\hat{\beta}_1))$$

    contains the true value of $\beta_1$.
  - $t^*$ is the upper critical value of a $t$ distribution with $n - 2$ degrees of freedom (df).

# Simulation Example

```r
simCI <- function() {
  f <- beta0 + beta1*x
  y <- f + rnorm(n,mean=0,sd=sd)
  ci <- confint(lm(y~x))
  ci["x",]
}
simCI() # Does it contain true value beta1 = 1?
```

```
##       2.5 %     97.5 %
## -0.3225214  1.0582357
```

```r
# Exercise: Write code to repeat NREPS times and count
# how many intervals include beta1.
simout <- replicate(NREPS,simCI())
sum(simout[1,]<= beta1 & simout[2,] >= beta1)
```

```
## [1] 955
```

# Advertising Example

```r
confint(afit)
```

```
##                  2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

- We say we are 95% confident that a $1000 increase in TV advertising is associated with an increase in sales of between 42 and 53 units.

# Hypothesis Tests

- The sampling distibution of the coefficients can also be used to derive tests of hypotheses about the parameters.
- Under the null hypothesis that the true $\beta_1$ is 0 (no association),

$$t = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} \sim t_{n-2}$$

- The usual alternative hypothesis is that $\beta_1 \neq 0$ (association). Then the p-value is the chance of $T > |t|$, where $T \sim t_{n-2}$.

# Advertising Example

```r
summary(afit)$coefficients
```

```
##                Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 7.03259355 0.457842940 15.36028 1.40630e-35
## TV          0.04753664 0.002690607 17.66763 1.46739e-42
```

- There is very good evidence that increasing TV advertising increases sales.

# Accuracy of the Model

- Two common measures of the ability of the model to explain variation in $Y$:
  1. The residual SE (RSE) $\sqrt{RSS/(n-2)}$, which is an estimator of $\sigma$.
  2. The $R^2 = \frac{TSS-RSS}{TSS}$, where $TSS$ is the total sum of squares

$$\sum_i (y_i - \bar{y})^2$$

    - $R^2$ is the proportion of variation in $Y$ explained by the regression on $X$.
- The $R^2$ is more commonly used as a goodness-of-fit measure.

# Advertising Example

```
summary(afit)
```

```
##
## Call:
## lm(formula = sales ~ TV, data = advert)
##
## 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
```

- TV advertising explains about 61% of the variation in sales.

# Residual Plots

- Residuals are the primary tool for checking model assumptions.
- For example, a plot of residuals versus fitted values can show evidence of
  - departures from linearity – look for nonlinear trends
  - departures from constant SD – look for funnel shapes
  - outliers – unusually large residuals

# Saving the Residuals and Fitted Values

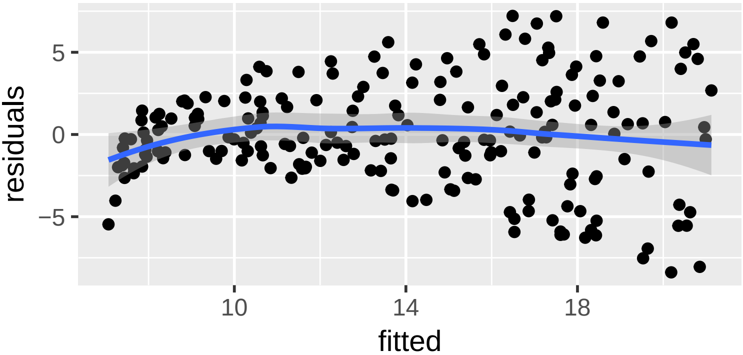
- Use the extractor functions `residuals()` and `fitted()`.

```
advertDiag <- data.frame(advert,residuals=residuals(afit),fitted=fitted(afit))
head(advertDiag)
```

```
##      TV radio newspaper sales  residuals    fitted
## 1 230.1  37.8      69.2  22.1  4.1292255 17.970775
## 2  44.5  39.3      45.1  10.4  1.2520260  9.147974
## 3  17.2  45.9      69.3   9.3  1.4497762  7.850224
## 4 151.5  41.3      58.5  18.5  4.2656054 14.234395
## 5 180.8  10.8      58.4  12.9 -2.7272181 15.627218
## 6   8.7  48.9      75.0   7.2 -0.2461623  7.446162
```

# Plotting Residuals vs. Fitted Values

```
ggplot(advertDiag,aes(x=fitted,y=residuals)) +
  geom_point() + geom_smooth()
```




- Some evidence of non-linearity on LHS of plot.
- Funnel from left to right.
- Consequences? Tendency to underestimate SE, which makes t-statistic too big and p-values too small.