THE UNIVERSITY OF TEXAS AT AUSTIN

McCOMBS BUSINESS ANALYTICS

Linear Regression Review

Book Chapter 3

Jared S. Murray
The University of Texas McCombs School of Business

Linear regression: Why?

Straight prediction questions:

- For how much will my house sell?
- How many runs per game will the Red Sox score this year?
- Will this person like that movie? (e.g., Netflix)

Explanation and understanding:

- What is the impact of getting an MBA on lifetime income?
- How do the returns of a mutual fund relate to the market?
- Does Walmart discriminate against women when setting salaries?

Example: Predicting House Prices

Problem:

- Predict market price based on observed characteristics

Solution:

- Look at property sales data where we know the price and some observed characteristics.
- Build a decision rule that predicts price as a function of the observed characteristics.

Predicting House Prices

What characteristics do we use?

We have to define the variables of interest and develop a specific quantitative measure of these variables

- Many factors or variables affect the price of a house
  - size
  - number of baths
  - garage
  - neighborhood
  - ...

Predicting House Prices

To keep things super simple, let's focus only on size.

The value that we seek to predict is called the dependent (or output) variable, and we denote this:

- Y, e.g. the price of the house (thousands of dollars)

The variable that we use to aid in prediction is the independent, explanatory, or input variable, and this is labelled

- X, e.g. the size of house (thousands of square feet)

How do we construct a prediction function?

Predicting House Prices

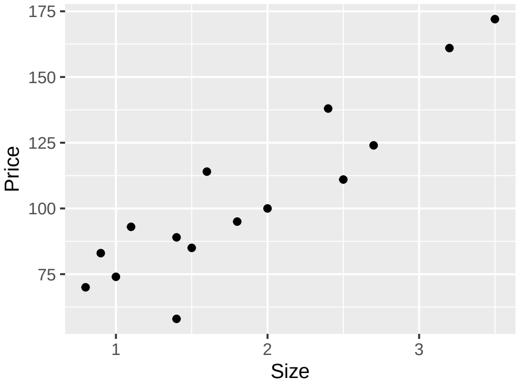


Table of Contents

Optimal linear prediction and ordinary least squares

The (OLS) Regression Table in R

R^2 and the "Variance Explained" by X

The SLR model

Multiple linear regression

Checking regression model assumptions

Transformations of X and/or Y

Optimal **Linear** Prediction

We'll restrict our attention to linear prediciton functions: $m(x) = b_0 + b_1 x$. The **mean squared error (MSE)** for a b_0, b_1 pair is

$$MSE(b_0, b_1) = E[(Y - [b_0 + b_1 X])^2]$$

Optimal **Linear** Prediction

The MSE takes a unique minimum at (β_0, β_1), where

$$\beta_0 = E(Y) - \beta_1 E(X)$$
$$\beta_1 = \frac{Cov(X, Y)}{Var(X)} = Cor(X, Y)\frac{SD(Y)}{SD(X)}$$

So the best linear prediction of Y when $X = x$ is

$$Y \approx E(Y) + \frac{Cov(X, Y)}{Var(X)}(X - E[X])$$

Optimal **Linear** Prediction

$$\beta_0 = E(Y) - \beta_1 E(X)$$
$$\beta_1 = \frac{Cov(X,Y)}{Var(X)} = Cor(X,Y)\frac{SD(Y)}{SD(X)}$$

$$Y \approx \beta_0 + \beta_1 X = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - E[X])$$

- The line of best fit passes through $(E[Y], E[X])$
- The slope of the line is the **correlation** times an adjustment factor $SD(Y)/SD(X)$ that accounts for the spread of X and Y (and their units of measurement)

Optimal **Linear** Prediction

$$\beta_0 = E(Y) - \beta_1 E(X)$$

$$\beta_1 = \frac{Cov(X,Y)}{Var(X)} = Cor(X,Y)\frac{SD(Y)}{SD(X)}$$

$$Y \approx \beta_0 + \beta_1 X = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - E[X])$$

- This should make clear the role of $Cor(X,Y)$ as a measure of **linear** dependence between X and Y – what happens to the slope if we measure price in yen and size in square meters?

Optimal **Linear** Prediction

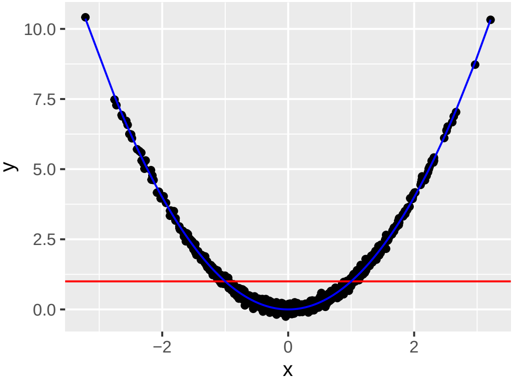
Some final notes about optimal linear prediction: To derive the optimal linear predictor, we don't have to say anything about:

- The joint distribution of (X, Y) or its marginals $P(Y)$ and $P(X)$ apart from existence of means/variances/covariances
- The distribution of prediction errors $Y - (\beta_0 + \beta_1 X)$
- Causal or temporal relationships between X and Y
- How close $\mu(x) = E(Y \mid X = x)$ is to $\beta_0 + \beta_1 x$ (i.e., whether the true μ is linear)

Optimal **Linear** Prediction

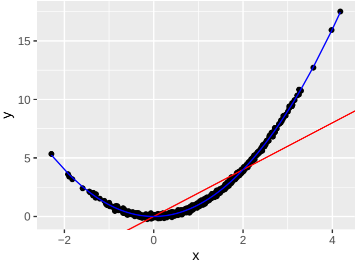
Two more important points:

1. The best linear prediction could be horrible. Suppose $X \sim N(0,1)$, and $Y \mid X \sim N(X^2, 0.1^2)$. The optimal linear predictor is $1 + 0 * x$ but....



Optimal **Linear** Prediction

2. The best linear prediction can change when $P(X)$ changes, even if the conditional distribution $P(Y \mid X)$ stays the same! Suppose now $X \sim N(1, 0.1^2)$, and $Y \mid X \sim N(X^2, 0.1^2)$ (unchanged). The optimal linear predictor is now $Y \approx 2 * X$



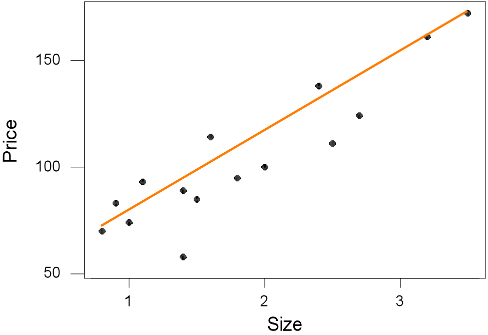
If the distribution of our inputs (house sizes) shifts, we should be wary of using past data to predict future outcomes (prices), unless we really believe the relationship is linear.

Estimating the Optimal Linear Predictor

Estimating the Linear Predictior

Based on the data, there appears to be a linear relationship between price and size:

As size goes up, price goes up.



The line shown was fit by the “eyeball” method.

Estimating the Optimal Linear Predictor

We want to use **data** to estimate the intercept and slope:

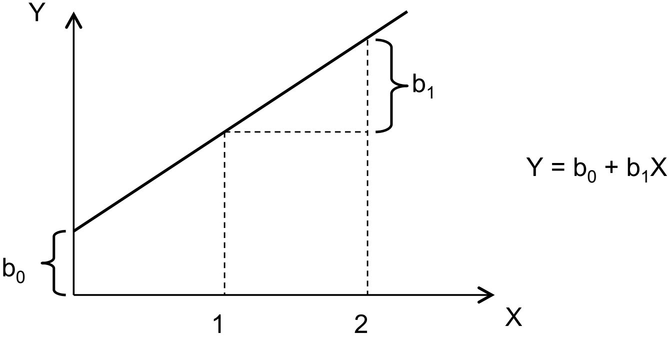
$$y = b_0 + b_1 x$$

Where b_0 is the intercept and b_1 is the slope.

The intercept value is in units of Y ($1,000).

The slope is in units of Y *per* units of X ($1,000/1,000 sq ft).

Estimating the Optimal Linear Predictor



Our “eyeball” line has $b_0 = 35$, $b_1 = 40$.

Estimation via plug-ins

Can we do better than the eyeball method? Yes!

We know that the optimal choice is β_0 and β_1, where

$$\beta_0 = E(Y) - \beta_1 E(X)$$
$$\beta_1 = \frac{Cov(X, Y)}{Var(X)}$$

But we don't know the means, variances, and covariance of X and Y, we just have a collection of data points (draws from $P(Y, X)$) $(y_1, x_1), (y_2, x_2), ..., (y_n, x_n)$

Estimation via plug-ins

$$\beta_0 = E(Y) - \beta_1 E(X)$$
$$\beta_1 = \frac{Cov(X,Y)}{Var(X)}$$

Substitute $\bar{y} = \sum_{i=1}^{n} y_i$ for $E(Y)$, $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})$ for $Var(X)$, etc.

Estimation by least squares

This is equivalent to minimizing the **sample mean squared error** (aka empirical MSE, in-sample MSE, sum of squared errors, sum of squared residuals...)

Recall: β_0, β_1 give the minimum true MSE:

$$(\beta_1, \beta_0) = \underset{(b_0,b_1)}{\arg\min}\, MSE(b_0, b_1) = \underset{(b_0,b_1)}{\arg\min}\, E[(Y - [b_0 + b_1 X])^2]$$

Estimation by least squares

Since we don't have access to the true $P(X, Y)$, we can approximate the MSE using our data as

$$E[(Y - [b_0 + b_1 X])^2] \approx \frac{1}{n} \sum_{i=1}^{n} (y_i - [b_0 + b_1 x_i])^2$$

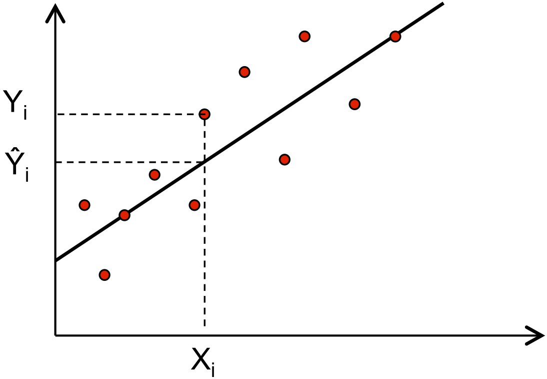
The **ordinary least squares (OLS) estimates** of β_0 and β_1 satisfy

$$(\hat{\beta}_1, \hat{\beta}_0) = \underset{(b_0, b_1)}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (y_i - [b_0 + b_1 x_i])^2$$

Let's build some intuition for this procedure...

Fitted values and residuals

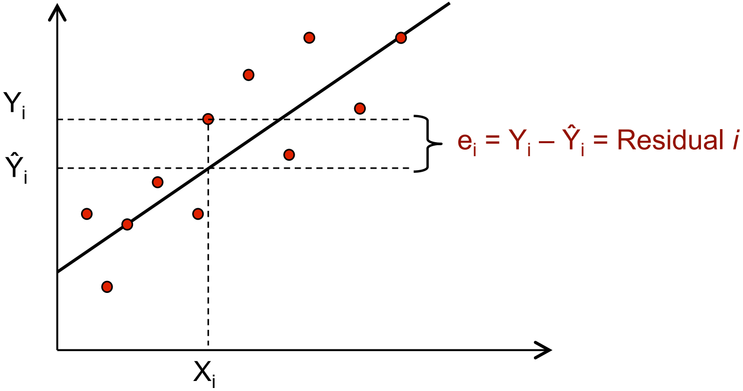
What is the "fitted value"?



The dots are the observed values and points on the line represents our fitted values given by $\hat{y}_i = b_0 + b_1 x_i$.

Fitted values and residuals

What is the “residual” for the *i*th observation?



We can write $y_i = \hat{y}_i + (y_i - \hat{y}_i) = \hat{y}_i + e_i$.

Least Squares

We want to minimize the size of all residuals:

- If they were all zero we would have a perfect line.
- Trade-off between moving closer to some points and at the same time moving away from other points.

This is what the OLS estimates do:

$$
\begin{aligned}
(\hat{\beta}_1, \hat{\beta}_0) &= \underset{(b_0, b_1)}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (y_i - [b_0 + b_1 x_i])^2 \\
&= \underset{(b_0, b_1)}{\arg\min} \sum_{i=1}^{n} e_i^2
\end{aligned}
$$

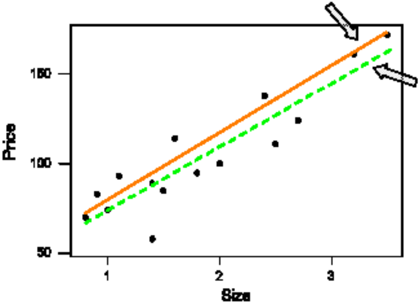
(where did the $\frac{1}{n}$ go?)

Least Squares

OLS gives a different line from our eyeball line:

- $\hat{\beta}_0 = 38.88$ and $\hat{\beta}_1 = 35.39$

Our line



LS line

The OLS Estimates

It turns out that minimizing the in-sample MSE gives

$$\hat{\beta}_1 = r_{xy} \times \frac{s_y}{s_x} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where,

- $\bar{X}$ and $\bar{Y}$ are the sample mean of X and Y
- $cor(x, y) = r_{xy}$ is the sample correlation
- s_x and s_y are the sample standard deviation of X and Y

The same as our plug-in estimator! These are the **ordinary least squares estimates** of β_0 and β_1.

Table of Contents

Optimal linear prediction and ordinary least squares

The (OLS) Regression Table in R

R^2 and the "Variance Explained" by X

The SLR model

Multiple linear regression

Checking regression model assumptions

Transformations of X and/or Y

Ordinary Least Squares in R

The `lm` command fits linear (regression) models

```
fit = lm(Price ~ Size, data = housing)
print(fit)

##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## Coefficients:
## (Intercept)         Size
##       38.88        35.39
```

```r
fit = lm(Price ~ Size, data = housing)
summary(fit)

##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.425  -8.618   0.575  10.766  18.498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.885      9.094   4.276 0.000903 ***
## Size          35.386      4.494   7.874 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 13 degrees of freedom
## Multiple R-squared:  0.8267,Adjusted R-squared:  0.8133
## F-statistic:    62 on 1 and 13 DF,  p-value: 2.66e-06
```

Ordinary Least Squares in R

We use **confidence intervals** to summarize uncertainty about parameter estimates.

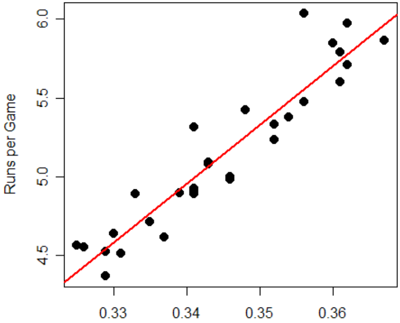
A confidence interval is **a set of plausible values for the parameter**.

Formally, a 95% confidence interval will capture the true value for the intercept/slope/other parameters in 95% of random samples from a given population.

```
confint(fit)

##                2.5 %   97.5 %
## (Intercept) 19.23850 58.53087
## Size        25.67709 45.09484
```

OLS: Summary



$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$$

- $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1$ is the slope
- We find them using *Least Squares*
- For a new value of the independent variable OBP (say x_{n+1}) we can predict the response y_{n+1} using the fitted line

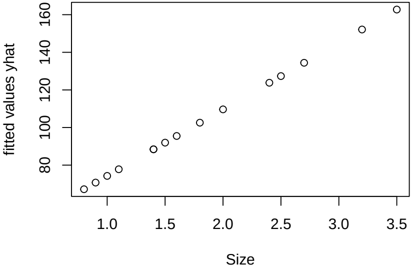
More on Least Squares

From now on, terms "fitted values" ($\hat{y}_i$) and "residuals" (e_i) refer to those obtained from the least squares line.

The fitted values and residuals have some special properties...

The Fitted Values and X

```
plot(predict(fit)~Size,
     data=housing,
     ylab="fitted values yhat")
```

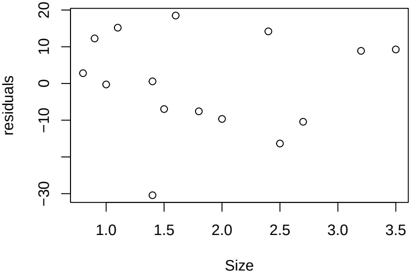


```
cor(predict(fit), housing$Size)

## [1] 1
```

The Residuals and X

```
plot(resid(fit)~Size,
     data=housing,
     ylab="residuals")
```



```
mean(resid(fit));  cor(resid(fit), housing$Size)

## [1] 2.368476e-16
## [1] 4.430566e-17
```

A Deeper Look at Least Squares Estimates

Least squares estimates have some special properties:

- The fitted values $\hat{Y}$ and x were **very** dependent
- The residuals $Y - \hat{Y}$ and x had no apparent relationship
- The residuals $Y - \hat{Y}$ had a sample mean of zero

What's going on?

Properties of Least Squares Estimates

In the housing data, we saw:

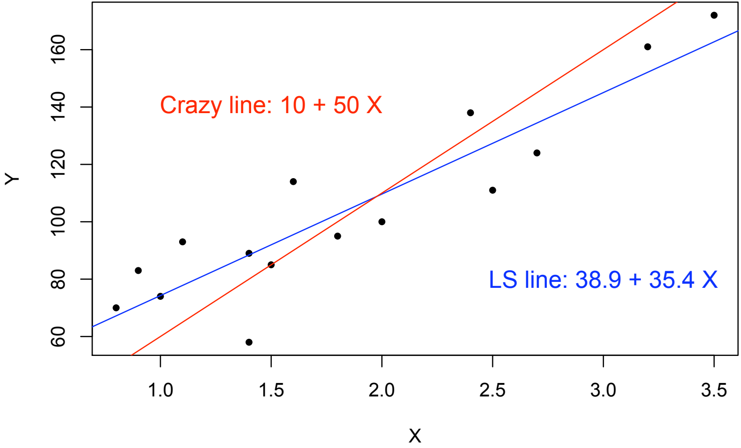
- $cor(\hat{Y}, x) = 1$ (a perfect linear relationship)
- $cor(e, x) = 0$ (no linear relationship)
- $mean(e) = 0$ (sample average of residuals is zero)

These facts are **always** true of OLS estimates

Why?

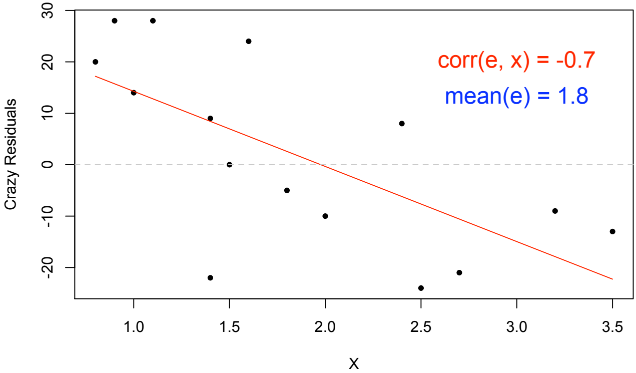
What is the intuition for the relationships between $\hat{y}$ and e and x?

Lets consider some "crazy" alternative line:



Fitted Values and Residuals

This is a bad fit! We are underestimating the value of small houses and overestimating the value of big houses, and our errors are positive on average



Clearly, we have left some predictive ability on the table!

Summary: LS is the best we can do!!

As long as the correlation between e and x is non-zero, we could always adjust our prediction rule to do better.

We need to exploit all of the predictive power in the x values and put this into $\hat{y}$, leaving no "*Xness*" in the residuals.

In summary: $y = \hat{y} + e$ where:

- $\hat{y}$ is "made from X" using a linear equation; $\mathrm{cor}(X, \hat{Y}) = \pm 1$.
- e has no **linear** relationship with X; $\mathrm{cor}(X, e) = 0$.
- On average (over the sample), our prediction error is zero: $\bar{e} = \sum_{i=1}^{n} e_i = 0$.

Table of Contents

Optimal linear prediction and ordinary least squares

The (OLS) Regression Table in R

R^2 and the "Variance Explained" by X

The SLR model

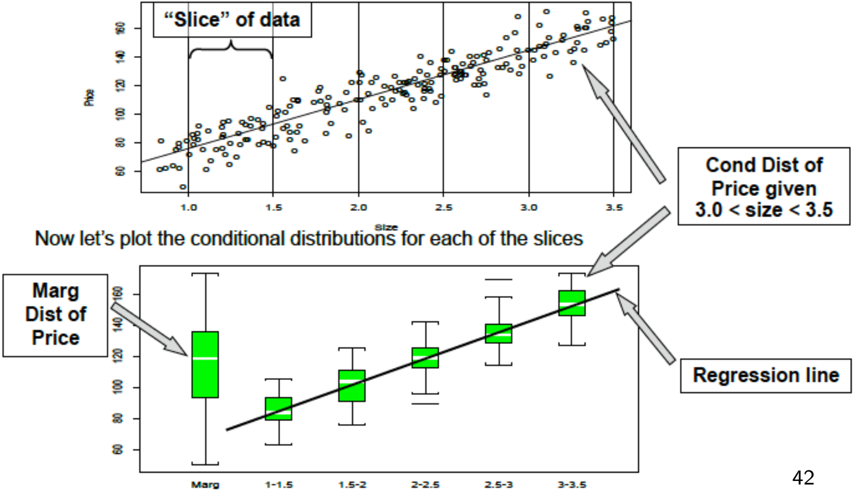
Multiple linear regression

Checking regression model assumptions

Transformations of X and/or Y

Decomposing the variability of outcomes

Using a “good” X reduces the unpredictability – i.e. the variability – in Y:



“Slice” of data

Cond Dist of Price given 3.0 < size < 3.5

Now let’s plot the conditional distributions for each of the slices

Marg Dist of Price

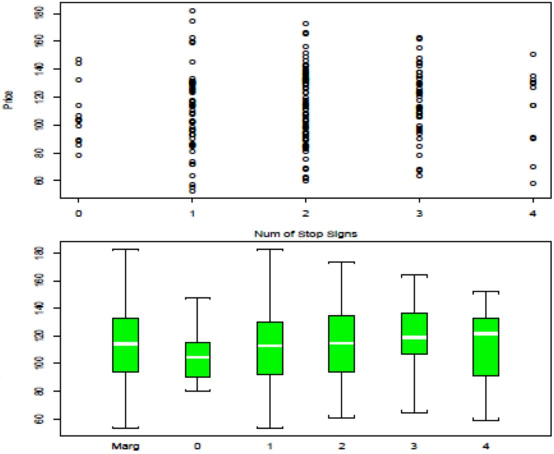
Regression line

Using a “bad” X doesn’t

When X has low predictive power, the story is different:

House price (Y) vs. the number of stop signs within a two block radius of a house (X).

See that in this case, the marginal and the Conditionals are not that different!



Decomposing the variance of Y

Go back to optimal linear prediction of Y from X when we know $P(X, Y)$

Remember:

- The variance of Y is the mean squared error when predicting Y by its **expected value**, i.e. without X.
- The variance of $Y - [\beta_0 + \beta_1 X]$ is the mean squared error when predicting Y from X using the optimal linear predictor

We can show that

$$Var(Y) = Var(Y - [\beta_0 + \beta_1 X]) + Var(\beta_0 + \beta_1 X)$$

What does that mean? So long as $\beta_1 \neq 0$, **the optimal linear prediction using X has lower MSE than not using X**

Decomposing the variance of Y

Remember, the optimal linear predictor isn't necessarily a good fit, and it doesn't assume or imply that X causes Y

- The decomposition is true even if the true relationship between X and Y is nonlinear
- The decomposition is true if X causes Y, or Y causes X, or something else causes both X and Y

It turns out that the OLS predictions estimates behave similarly, with sample quantities replacing true quantities.

Decomposing the Variance

Remember that $Y = \hat{Y} + e$

Since $\hat{Y}$ and e are uncorrelated, i.e. $\mathrm{cor}(\hat{Y}, e) = 0$,

$$\begin{aligned}
\mathrm{var}(Y) &= \mathrm{var}(\hat{Y} + e) = \mathrm{var}(\hat{Y}) + \mathrm{var}(e) \\
\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} &= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}{n-1} + \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n-1}
\end{aligned}$$

Given that $\bar{e} = 0$, and the sample mean of the fitted values $\bar{\hat{Y}} = \bar{Y}$ (why?) we get to write:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2$$

Decomposing the Variance

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{\boxed{\begin{array}{c}\textbf{Total Sum of}\\ \textbf{Squares}\\ \textbf{SST}\end{array}}}=\underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{\boxed{\begin{array}{c}\textbf{Regression SS}\\ \textbf{SSR}\end{array}}}+\underbrace{\sum_{i=1}^{n}e_i^{\,2}}_{\boxed{\begin{array}{c}\textbf{Error SS}\\ \textbf{SSE}\end{array}}}$$

- SST is measuring *total variation in Y/total error in Y using the simplest prediction $\bar{Y}$ – i.e., no info about X*
- SSR is measuring *predictable (via our regression model) variation in Y – how much our predictions change after accounting for linear effects of X*
- SSE is measuring *left over, unpredictable variation in Y*

Decomposing the Variance

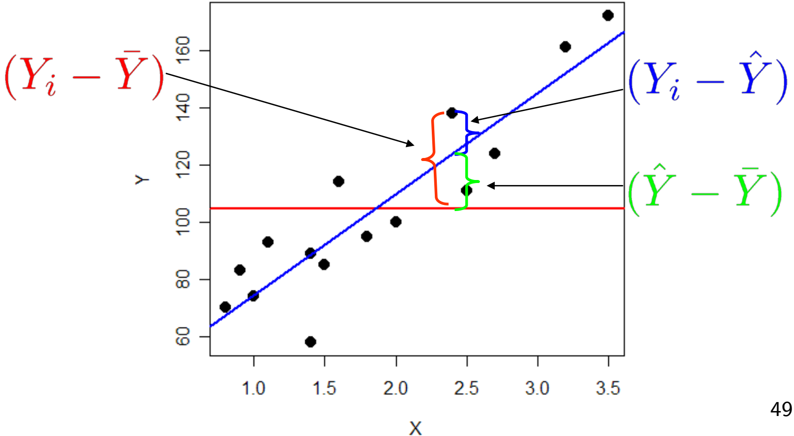
$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{\boxed{\begin{array}{c}\textbf{Total Sum of}\\ \textbf{Squares}\\ \textbf{SST}\end{array}}}=\underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{\boxed{\begin{array}{c}\textbf{Regression SS}\\ \textbf{SSR}\end{array}}}+\underbrace{\sum_{i=1}^{n}e_i^{\,2}}_{\boxed{\begin{array}{c}\textbf{Error SS}\\ \textbf{SSE}\end{array}}}$$

Things to note:

- SST is fixed, so as SSR increases, SSE (the total error in our predictions) goes down.
- SSR describes variation that's predictable by a **linear** equation of X. We could get better SSR (and lower SSE) with **nonlinear** functions of X, but we have to be careful – more soon.

Decomposing the Variance

$$
\begin{aligned}
(Y_i - \bar{Y}) &= \hat{Y}_i + e_i - \bar{Y} \\
&= (\hat{Y}_i - \bar{Y}) + e_i
\end{aligned}
$$



The Coefficient of Determination R^2

The coefficient of determination, denoted by R^2, measures how well the fitted values $\hat{Y}$ follow Y:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}$$

- R^2 is often called the proportion of variance in Y that is “explained” by the regression line (in the mathematical – not scientific – sense!): $R^2 = 1 - Var(e)/Var(Y)$
- $0 < R^2 < 1$
- For simple linear regression, $R^2 = r_{xy}^2$. Similar caveats to sample correlation apply!

The Coefficient of Determination R^2

For the optimal linear predictor we also have

$$R^2 = 1 - \frac{Var(Y - [\beta_0 + \beta_1 X])}{Var(Y)} = Cor(X, Y)^2$$

This is a little easier to show, but the idea is the same.

Explanations and predictions

A better way to think about R^2 is as the **proportion of variability – i.e. unpredictablility – in Y that becomes predictable when using X in a linear regression model.**

Remember:

$$\begin{aligned} R^2 &= 1 - Var(e)/Var(Y) \\ &= 1 - \frac{\text{average squared prediction error of linear model with } X}{\text{average squared prediction error ignoring } X} \end{aligned}$$

Explanations and predictions

A better way to think about R^2 is as the **proportion of variability – i.e. unpredictablility – in Y that becomes predictable when using X in a linear regression model.**

R^2 does not tell you:

- Whether there is/is not any causal relationship between X and Y (Question: What is the R^2 from regressing X on Y)?
- Whether your regression model is a reasonable approximation of reality
- Whether your model predicts well on new data or generalizes well outside your sample

R^2 for the Housing Data

```
summary(fit)

##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.425  -8.618   0.575  10.766  18.498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.885      9.094   4.276 0.000903 ***
## Size          35.386      4.494   7.874 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 13 degrees of freedom
## Multiple R-squared:  0.8267,Adjusted R-squared:  0.8133
```

R^2 for the Housing Data

```
summary(fit)

##
## Call:
## lm(formula = Price ~ Size, data = housing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.425  -8.618   0.575  10.766  18.498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   38.885      9.094   4.276 0.000903 ***
## Size          35.386      4.494   7.874 2.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.14 on 13 degrees of freedom
## Multiple R-squared:  0.8267, Adjusted R-squared:  0.8133
## F-statistic:    62 on 1 and 13 DF,  p-value: 2.66e-06
```

R^2 for the Housing Data

```
anova(fit)

## Analysis of Variance Table
##
## Response: Price
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## Size       1 12393.1 12393.1  61.998 2.66e-06 ***
## Residuals 13  2598.6   199.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

$$R^2 = \frac{SSR}{SST} = \frac{12393.1}{2598.6 + 12393.1} = 0.8267$$

Table of Contents

Optimal linear prediction and ordinary least squares

The (OLS) Regression Table in R

R^2 and the “Variance Explained” by X

The SLR model

Multiple linear regression

Checking regression model assumptions

Transformations of X and/or Y

Overview

So far we've derived the **optimal linear predictor** of Y using X when we know $P(X, Y)$, and how to estimate it from data if we don't know $P(X, Y)$

We made almost no assumptions to do so!

But to really put regression to work, and understand its properties, it will help to introduce some more modeling assumptions...

The Simple Linear Regression Model

Remember that a statistical or probability model is simply a set of assumptions on the probability distribution that did or could have generated our data.

Our first simple linear regression model is:

1. The distribution of X is arbitrary (and perhaps X is even non-random).
2. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some constants ("coefficients", "parameters") β_0 and β_1, and some random noise variable ϵ. (the "error term")
3. $E(\epsilon | X = x) = 0$ (no matter what x is), $Var(\epsilon | X = x) = \sigma^2$ (no matter what x is).
4. ϵ is uncorrelated across observations.

The Simple Linear Regression Model

Taking these in turn:

1. The distribution of X is arbitrary (and perhaps X is even non-random).

In the case that X is non-random, we may assume it can still be *described* by a probability distribution when we need to.

This is basically a non-assumption.

The Simple Linear Regression Model

2. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some constants (“coefficients”, “parameters”) β_0 and β_1, and some random noise variable ϵ.

We’re assuming that there is truly a linear relationship between X and Y. This is a big leap!

ϵ represents unobserved factors influencing Y that can be *treated* as random noise when we don’t know them, as well as more familiar forms of “random noise” (like measurement error)

The Simple Linear Regression Model

3. $E(\epsilon|X = x) = 0$ **(no matter what** x **is),** $Var(\epsilon|X = x) = \sigma^2$ **(no matter what** x **is).**

There are two assumptions here:

$E(\epsilon|X = x) = 0$ at any x is the **mean independence** (between X and ϵ) assumption. It means that the expected value of the idiosyncratic component ϵ can't depend on the value of X, i.e. must be constant. If it's constant it must be zero, so that the "level" is controlled by β_0.

$Var(\epsilon|X = x) = \sigma^2$ means that we have **homoskedastic errors** – their spread doesn't depend on X, or anything else.

The Simple Linear Regression Model

4. ϵ is uncorrelated across observations.

The assumption says that the errors – the part of Y that doesn't depend on X in the model – are unrelated. This could be violated, for example, if our observations came from a time series and X didn't "soak up" the time trend.

Sampling Distributions Under the SLR Model

The extra assumptions in the SLR model are let us make statements about the *sampling distribution* of $\hat{\beta}_0, \hat{\beta}_1$ and get confidence intervals and p-values.

In particular: The OLS estimates are unbiased, i.e.

$$E(\hat{\beta}_0) = \beta_0,\ E(\hat{\beta}_1) = \beta_1$$

What about their sampling variability (standard errors?)

Standard Errors Under the SLR Model

What factors do we think should influence the variability of the *OLS slope* $\hat{\beta}_1$ over repeated sampling, assuming our model holds?

Recall: $Y = \beta_0 + \beta_1 X + \epsilon$, $Var(\epsilon \mid X) = Var(\epsilon) = \sigma^2$

Standard Errors Under the SLR Model

What factors do we think should influence the variability of the *OLS slope* $\hat{\beta}_1$ over repeated sampling, assuming our model holds?

Recall: $Y = \beta_0 + \beta_1 X + \epsilon$, $Var(\epsilon \mid X) = Var(\epsilon) = \sigma^2$

$$s^2_{\hat{\beta}_1} = Var(\hat{\beta}_1) = \frac{\sigma^2}{n s_x^2}$$

- σ^2 is the remaining variability in Y after accounting for X (error variance)
- n is the sample size
- $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sample variance of the predictor (using n over $n-1$ is much simpler here!)

Sampling Distributions Under the SLR Model

What factors do we think should influence the variability of the *OLS intercept* $\hat{\beta}_0$ over repeated sampling, assuming our model holds?

Recall: $Y = \beta_0 + \beta_1 X + \epsilon$, $Var(\epsilon \mid X) = Var(\epsilon) = \sigma^2$

Sampling Distributions Under the SLR Model

What factors do we think should influence the variability of the *OLS intercept* $\hat{\beta}_0$ over repeated sampling, assuming our model holds?

Recall: $Y = \beta_0 + \beta_1 X + \epsilon$, $Var(\epsilon \mid X) = Var(\epsilon) = \sigma^2$

$$s^2_{\hat{\beta}_1} = Var(\hat{\beta}_0) = \frac{\sigma^2}{n}\left(1 + \left(\frac{\bar{x}}{s_x}\right)^2\right)$$

- σ^2 is the remaining variability in Y after accounting for X (error variance)
- n is the sample size
- $|\bar{x}|/s_x$ is the standardized distance between zero and $\bar{x}$

Sampling Distributions Under the SLR Model

What factors do we think should influence the variability of the *OLS prediction* $\hat{\beta}_0 + \hat{\beta}_1 x$ over repeated sampling, assuming our model holds?

Recall: $Y = \beta_0 + \beta_1 X + \epsilon$, $Var(\epsilon \mid X) = Var(\epsilon) = \sigma^2$

Sampling Distributions Under the SLR Model

What factors do we think should influence the variability of the *OLS prediction* $\hat{\beta}_0 + \hat{\beta}_1 x$ over repeated sampling, assuming our model holds?

Recall: $Y = \beta_0 + \beta_1 X + \epsilon,\ Var(\epsilon \mid X) = Var(\epsilon) = \sigma^2$

$$s^2_{\hat{y}(x)} = Var(\hat{\beta}_0 + \hat{\beta}_1 x) = \frac{\sigma^2}{n}\left(1 + \left(\frac{x - \bar{x}}{s_x}\right)^2\right)$$

- σ^2 is the remaining variability in Y after accounting for X (error variance)
- n is the sample size
- $|x_i - \bar{x}|/s_x$ is the standardized distance between x and $\bar{x}$

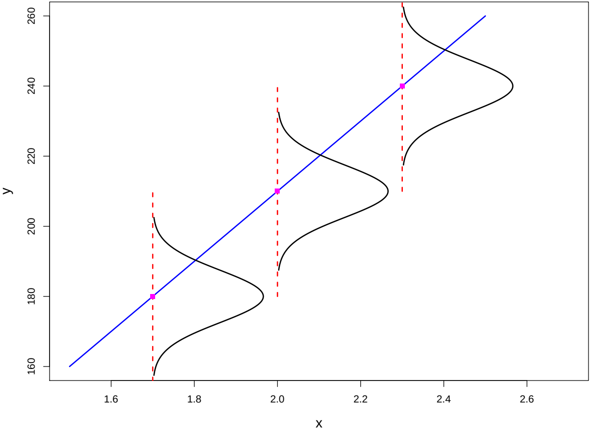
The SLR model with Gaussian errors

We can get even further if we're willing to make stronger modeling assumptions:

1. The distribution of X is arbitrary (and perhaps X is even non-random).
2. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some constants ("coefficients", "parameters") β_0 and β_1, and some random noise variable ϵ. (the "error term")
3. $\epsilon \sim N(0, \sigma^2)$, is independent of X, and is independent across observations

All our previous assumptions, plus assuming that the errors have a normal distribution

The Simple Linear Regression Model



The conditional distribution for Y given $X = x$ is normal (why?):

$$(Y|X = x) \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2).$$

The SLR model with Gaussian errors: Prediction uncertainty

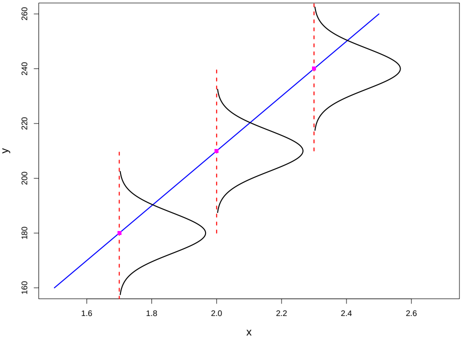
Imagine we're trying to make a prediction of the sale price of a 2000 sqft house. Our best guess using our linear model is

$$\hat{\beta}_0 + \hat{\beta}_1 2$$

We know its sampling distribution, which measures our uncertainty about the true optimal linear prediction. Is that all the uncertainty we have in the actual sale price for the single house?

The Simple Linear Regression Model: Prediction uncertainty

We have to account for uncertainty about $\beta_0 + \beta_1 x$ AND the random term ϵ



$$(Y|X = x) \sim \mathrm{N}(\beta_0 + \beta_1 x, \sigma^2).$$

The Simple Linear Regression Model: Prediction uncertainty

95% prediction interval for the sale price of **a single** 2000sqft house:

```r
predict(fit, newdata=data.frame(Size=2),
        interval = 'prediction', level = 0.95)

##        fit      lwr      upr
## 1 109.6566 78.07862 141.2346
```

Includes uncertainty about the line *and* how far an individual house is from it’s predicted value.

The Simple Linear Regression Model: Prediction uncertainty

95% confidence interval for the **average** sale price across **all** 2000sqft houses:

```
predict(fit, newdata=data.frame(Size=2),
        interval = 'confidence', level = 0.95)

##        fit      lwr      upr
## 1 109.6566 101.6426 117.6706
```

Includes uncertainty about the line only.

Table of Contents

Optimal linear prediction and ordinary least squares

The (OLS) Regression Table in R

R^2 and the "Variance Explained" by X

The SLR model

Multiple linear regression

Checking regression model assumptions

Transformations of X and/or Y

Multiple Linear Regression

In multiple linear regression we utilize several predictors in a linear model:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \epsilon$$

Instead of a regression *line*, we now estimate a regression *plane*.

Multiple Linear Regression

If $p = 2$, we can plot the regression surface in 3D.

Consider sales of a product as predicted by price of this product (P1) and the price of a competing product (P2).

$$Sales = \beta_0 + \beta_1 P1 + \beta_2 P2 + \epsilon$$

Multiple Linear Regression

Most of what we learned via SLR carries over directly:

- We fit the model via OLS; definitions of the OLS coefficients $\hat{\beta}_j$, fitted values $\hat{y}_i = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_j x_{ij} = \mathbf{x_i}'\beta$, and residuals $y_i - \hat{y}_i$ stay the same.
- The fitted values and residuals satisfy the same properties as in SLR: $cor(\hat{y}, e) = 0$, $cor(x_j, e) = 0$ for all covariates in the model, $mean(e) = 0$.
- $R^2 = 1 - \frac{var(e)}{var(y)} = cor^2(y, \hat{y})$ measures the proportion of variability in y predictable via a linear model using these covariates

Multiple Linear Regression Model

The MLR model with Gaussian errors is essentially identical to SLR:

- There are p quantitative predictor variables, $X_1, X_2, \ldots X_p$. X without a subscript will refer to the vector of all of these taken together.
- There is a single response variable Y.
- The variables are related through

  $$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \epsilon,$$

  for some constants (coefficients) $\beta_0, \beta_1, \ldots \beta_p$.
- The noise variables ϵ are independent and identically distributed $N(0, \sigma^2)$, independent of $\mathbf{X}$.

Multiple Linear Regression Model

Most of what we learned about the SLR model carries over directly:

- So long as the error terms are iid and uncorrelated with X, the OLS estimates are unbiased; their sampling variance is a little different.
- If the error terms are also Gaussian, then the sampling distribution of the OLS estimates is Gaussian. Otherwise, their sampling distribution will eventually look Gaussian by the CLT
- We can use these facts for confidence intervals and hypothesis tests. The bootstrap works in the same fashion as well.

Understanding Multiple Linear Regression

The biggest difference in MLR is the interpretation of the coefficients. Note that:

$$\beta_j = \frac{\partial E[Y|X_1, \ldots, X_p]}{\partial X_j}$$

measures a *partial* effect of X_i on the predicted value of Y. It's partial because we're assuming that **all the other covariates are held constant**

But what does that mean?

Case study: Spending more to get less?

There is a long-running debate about the relationship between spending on public education – specifically high school – and educational outcomes.

In 1993, columnist George Will noted that:

- Of the 10 states with **lowest** per pupil spending, 4 are among the 10 states with top average SAT scores
- Of the 10 states with **highest** per pupil spending, only 1 is among the 10 states with top average SAT scores
- NJ has the highest per pupil spending and only the 39th highest SAT scores

Will concludes: "The public education lobby's crumbling last line of defense is the miseducation of the public."

Spending more to get less?

Is he right? Let's look at the data (`school_expenditures.R`)

Spending more to get less?: Summary

- Summary:
  1. If we don't include the SAT participation rate in the model, expenditures have a negative **overall** relationship with outcomes (SAT scores)
  2. But we know that in states with low participation rates, the participants are, on average, more likely to be high performers applying to elite out-of-state schools! In addition, many higher spending states are on the east coast, where SAT scores were preffered to ACT, leading to high participation rates

Spending more to get less?: Summary

- Summary:
  1. If we include the SAT participation rate in the model, expenditures have a positive **partial** relationship with outcomes (SAT scores)
  2. MLR gives us a tool for comparing like-to-like: The regression coefficient on expenditures in the model including both expenditures and participation is the expected change in average SAT score for a $1000 increase in per-pupil expenditures, **adjusting for/holding constant the SAT participation rate**

Interpreting regression coefficients

In general: Our interpretation of regression coefficients is...

Holding all other variables in the model constant, β_j is the average/predicted/expected change in Y per unit change in X_j.

Table of Contents

Optimal linear prediction and ordinary least squares

The (OLS) Regression Table in R

R^2 and the “Variance Explained” by X

The SLR model

Multiple linear regression

Checking regression model assumptions

Transformations of X and/or Y

Diagnostics for the linear model

Modeling assumptions include:

1. The distribution of X is arbitrary (and perhaps X is even non-random).
2. If $X = x$, then $Y = \beta_0 + \beta_1 x + \epsilon$, for some constants ("coefficients", "parameters") β_0 and β_1, and some random noise variable ϵ. (the "error term")
3. $\epsilon \sim N(0, \sigma^2)$, is independent of X, and is independent across observations

All of these assumptions can be consequential, some more than others. Before relying on the model, we should check its assumptions!

Most of our assumptions are (or can be) stated about the error term, so the residuals will be our friend here...

Properties of the residuals when the model fits well

1. The residuals should have expectation zero, conditional on x, $E(e_i|X = x) = 0$.
2. The (standardized) residuals should show a constant variance, unchanging with x.
3. The residuals can't be completely uncorrelated with each other, but the correlation should be extremely weak, and grow negligible as $n \to \infty$.
4. If the noise is Gaussian, the residuals should also be Gaussian.

We can use graphical checks for all of these.

Let's see an example

`diagnostics.R`

Table of Contents

Optimal linear prediction and ordinary least squares

The (OLS) Regression Table in R

R^2 and the "Variance Explained" by X

The SLR model

Multiple linear regression

Checking regression model assumptions

Transformations of X and/or Y

Transformations in regression models

Sometimes it makes more sense to model $g(Y)$ or $f(X)$ instead of the variables X or Y as they arrived to us.

In many regression classes transformations are pitched as a “fix” for failed diagnostics

While a transformation leading to model fit will improve diagnostics, you should always consider the implications for interpreting the model.

Transforming the predictor

$$Y = \beta_0 + \beta_1 f(X) + \epsilon$$

Essentially we just define a new covariate $f(X)$ and proceed

How does the interpretation of β_0 and β_1 change?

- $\beta_0 = E[Y \mid f(X) = 0]$ (the intercept when $f(x)$ is on the $x-$axis)
- When $f(X)$ increases by one unit the predicted value for Y increases by β_1
- $\frac{dE[Y|X=x]}{dx} = \beta_1 f'(x)$ (when the derivative exists)

Example: Log transforms of X

- $\log X = 0$ means $X = 1$, so $\beta_0 = \mathbb{E}[Y \mid X = 1]$.
- A k unit change in $\log x$ means multiplying x by e^k :

  $$k + \log x = \log e^k + \log x = \log xe^k$$

  Hence, β_1 is the expected difference in Y for an e-fold change in X.
- The slope of $E[Y]$ with respect to X decreases in x :

  $$\frac{dE[Y \mid X = x]}{dx} = \frac{\beta_1}{x}$$

Transforming the response

$$g(Y) = \beta_0 + \beta_1 X + \epsilon$$

Essentially we just define a new response variable $g(Y)$ and proceed

Your textbook explores in some depth how to interpret this model in terms of the original Y. It is... challenging.

Except in very special cases (log transforms again), just interpret the model in terms of the transformed response itself (instead of Y)

Example: Log transforms of Y

$$\log Y = \beta_0 + \beta_1 X + \epsilon$$

On the original scale,

$$Y = e^{\beta_0 + \beta_1 X + \epsilon} = e^{\beta_0} e^{\beta_1 x} e^{\epsilon}$$

This gives us a *multiplicative* model!

Note: Even though $E(e^{\epsilon}) \neq 0$, it is still common to use $e^{\beta_0} e^{\beta_1 x}$ as our prediction for Y – i.e., predict $log(Y)$ and then exponentiate

Example: Log transforms of Y

$$Y = e^{\beta_0+\beta_1 X+\epsilon} = e^{\beta_0}e^{\beta_1 x}e^{\epsilon}$$

- e^{β_0} is the median of Y when $X = 0$ (Why?)
- A one unit increase in X increases the predicted value of Y by a *multiplicative factor* of e_1^{β}

$$e^{\beta_0}e^{\beta_1(x+1)} = [e^{\beta_0}e^{\beta_1 x}]e^{\beta_1}$$

Transforming the predictor and the response

$$g(Y) = \beta_0 + \beta_1 f(X) + \epsilon$$

Basically we have the all the problems and considerations from the previous cases. Unless g changes the response into something interpretable, this is going to be a difficult model to understand

We have one more interesting special case...

Example: Log transforms of X and Y

$$\log Y = \beta_0 + \beta_1 \log X + \epsilon$$

On the original scale,

$$Y = e^{\beta_0} X^{\beta_1} e^{\epsilon}$$

Again,common to use $e^{\beta_0} X_1^{\beta}$ as our prediction for Y – i.e., predict $log(Y)$ and then exponentiate

- e^{β_0} is the median of Y when $X = 1$ (Why?)
- β_1 is the slope of $\log Y$ against $\log X$. But what does that mean?

Example: Log transforms of X and Y

$$\log Y = \beta_0 + \beta_1 \log X + \epsilon$$

On the original scale,

$$Y = e^{\beta_0} X^{\beta_1} e^{\epsilon}$$

Again,common to use $e^{\beta_0} X_1^{\beta}$ as our prediction for Y – i.e., predict $log(Y)$ and then exponentiate

- e^{β_0} is the median of Y when $X = 1$ (Why?)
- β_1 is the slope of $\log Y$ against $\log X$. But what does that mean?

Example: Log transforms of X and Y

β_1 is the *elasticity*:

$$\frac{\frac{d\hat{y}}{dx}}{\frac{\hat{y}}{x}} = \frac{y_0\beta_1 x^{\beta_1 - 1}}{\frac{y_0 x^{\beta_1}}{x}} = \beta_1 = \frac{\frac{d\hat{y}}{dx}}{\frac{\hat{y}}{x}}$$

Think about $\frac{\frac{d\hat{y}}{dx}}{\frac{\hat{y}}{x}}$ as

$$\frac{\frac{d\hat{y}}{\hat{y}}}{\frac{dx}{x}}$$

Roughly, the percentage change in Y per *percentage change* in X.
Even more roughly: For a 1 percent increase in X, there is a $\approx \beta_1$ percent increase in the predicted value of Y.