LINEAR REGRESSION: INFERENCE

Connor K. Brubaker

Department of Statistics
Texas A&M University

Linear Model

Suppose we have data $(X_1, Y_1), \ldots, (X_n, Y_n)$. The simple linear regression model states that

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ is the error term.

- β_0 is the intercept parameter
- β_1 is the slope parameter

Assumptions of the Linear Model

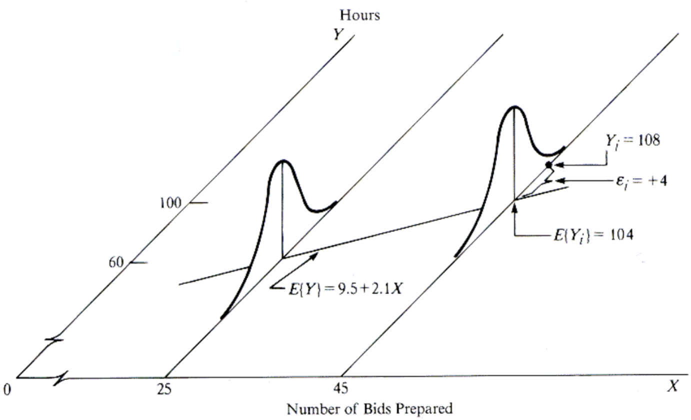
The simple linear model makes four assumptions:

1. X and Y are linearly related,
2. the errors $\varepsilon_1, \ldots, \varepsilon_n$ are independent of each other,
3. the errors $\varepsilon_1, \ldots, \varepsilon_n$ have a common variance σ^2, and
4. the errors $\varepsilon_1, \ldots, \varepsilon_n$ are normally distributed with a mean of 0 and variance σ^2.

For now, assume these are satisfied. We will revisit checking these assumptions (known as **model validation**) in the following lecture.

Assumptions of the Linear Model

FIGURE 1.6 Illustration of Simple Linear Regression Model (1.1).



Source: https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression4.html

LEAST SQUARES ESTIMATORS

Recall that the least squares estimators are

$$\hat{\beta}_1 = \frac{SXY}{SXX} \text{ and } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

where

$$SXY = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \text{ and } SXX = \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

Estimation of the Error Variance σ^2

Using the least squares estimates, the ith error is estimated with

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$$

The variance of the errors σ^2 is estimated using

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \frac{RSS}{n-1}$$

Inference for the Slope Parameter β_1

As long as the assumptions of the simple linear model are satisfied, it can be shown that

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{SXX}\right)$$

where β_1 is the true (unknown) slope parameter and σ^2 is the variance of the error terms. Therefore, the standard error of the slope is

$$\mathrm{SE}(\hat{\beta}_1) = \frac{\sigma}{\sqrt{SXX}}$$

where σ is estimated using $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$.

Confidence Interval for the Slope

A $(1 - \alpha) \times 100\%$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm t^*_{\alpha/2,n-2} \times \mathrm{SE}(\hat{\beta}_1) = \hat{\beta}_1 \pm t^*_{\alpha/2,n-2} \times \frac{\hat{\sigma}}{\sqrt{SXX}}$$

The degrees of freedom is $n - 2$ since we are estimating 2 parameters (the slope and the intercept).

Hypothesis Testing for the Slope

Often we want to test whether a significant linear relationship exists between X and Y. If no relationship exists, then

$$\mathrm{Cov}(X, Y) = 0$$

and consequently $\beta_1 = 0$. Therefore, we want to test the hypotheses

$$H_0 : \beta_1 = 0 \ \text{ versus } \ H_A : \beta_1 \neq 0$$

This is the test `R` performs by default.

Hypothesis Testing for the Slope

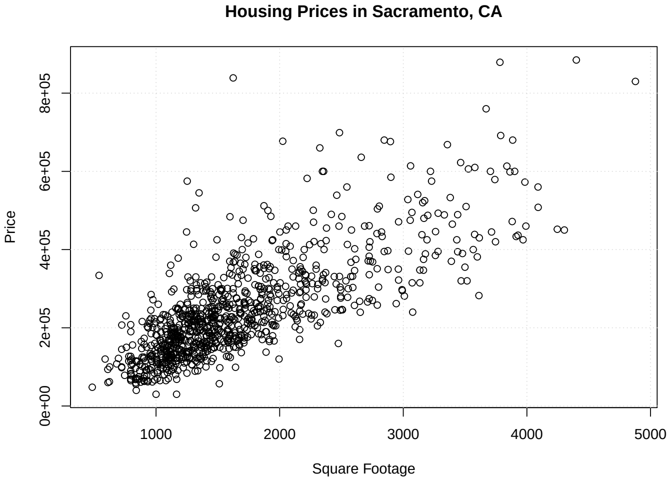
Let β_1^0 be the null value of the slope (often 0).

The test statistic for the hypothesis test is

$$T_{obs} = \frac{\hat{\beta}_1 - \beta_1^0}{\mathrm{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1^0}{\hat{\sigma}/\sqrt{SXX}} \sim t_{n-2}$$

You would use the t distribution with $n - 2$ degrees of freedom to find critical values and determine p-values.

Housing Prices



Reading R Output

```
> summary(model)

Call:
lm(formula = price ~ sqft, data = Sacramento)

Residuals:
    Min      1Q  Median      3Q     Max
-231889  -54717  -11822   38993  600141

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13859.393   6948.714   1.995   0.0464 *
sqft          138.546      3.796  36.495   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84130 on 930 degrees of freedom
Multiple R-squared:  0.5888,    Adjusted R-squared:  0.5884
F-statistic:  1332 on 1 and 930 DF,  p-value: < 2.2e-16
```

Summary statistics for the residuals

$\hat{\sigma}$

READING R OUTPUT

```
> summary(model)

Call:
lm(formula = price ~ sqft, data = Sacramento)

Residuals:
    Min      1Q  Median      3Q     Max
-231889  -54717  -11822   38993  600141

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13859.393   6948.714   1.995   0.0464 *
sqft          138.546      3.796  36.495   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 84130 on 930 degrees of freedom
Multiple R-squared:  0.5888,	Adjusted R-squared:  0.5884
F-statistic:  1332 on 1 and 930 DF,  p-value: < 2.2e-16
```

T_{obs}

$SE(\hat{\beta}_1)$

p-value

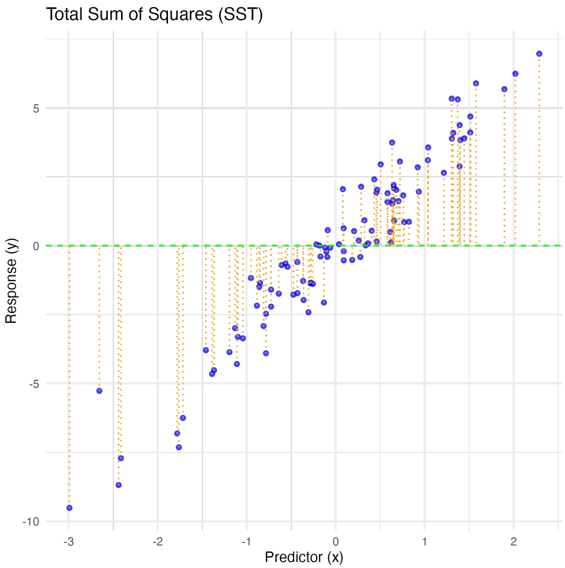
Analysis of Variance

The analysis of variance for a regression model allows us to determine how much of the variability in the data is captured by the model, i.e., how effective the model is. Begin by defining

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \geq 0$$

SST is the total sum of squares and represents the total variability in the response.

Total Sum of Squares



The analysis of variance *decomposes* the total variability into two terms:

$$SST = SSR + SSE$$

Therefore, $0 \leq SSR \leq SST$ and $0 \leq SSE \leq SST$.

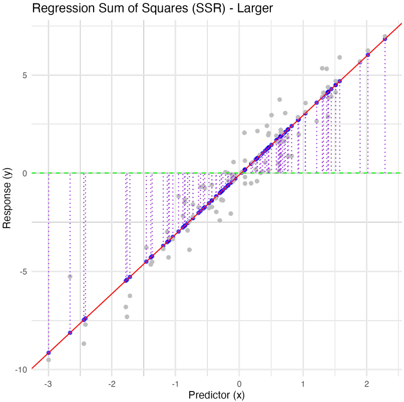
ANALYSIS OF VARIANCE

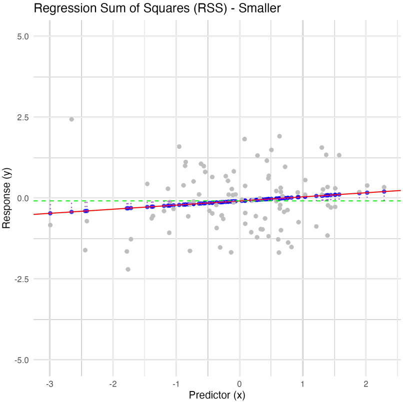
We will determine how effective the model is by determining what proportion of SST is captured by the model. The **regression** sum of squares is

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

- SSR is large if fitted values are far from the mean response
- SSR is small if the fitted values are all near the mean response

Regression Sum of Squares (SSR)





Residual Sum of Squares

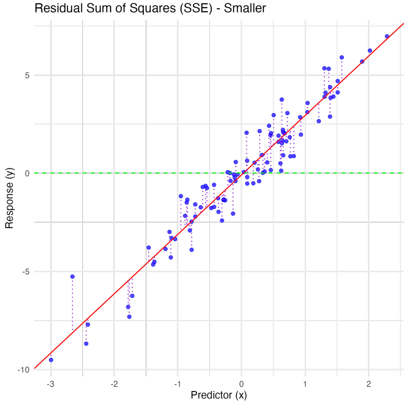
We've already seen the residual sum of squares,

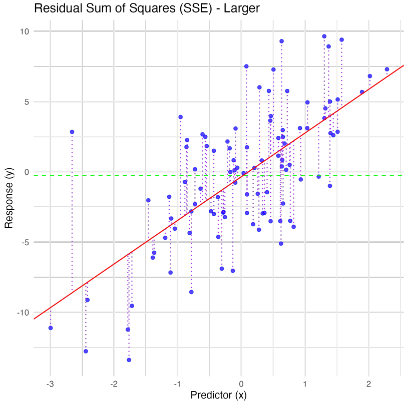
$$RSS \text{ or } SSE = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

which is also called the sum of squared errors *SSE*.

- Large *RSS* means lots of scatter around the regression line.

Residual Sum of Squares (SSE)





Coefficient of Determination R^2

The **coefficient of determination** is

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

- R^2 is the square of the correlation between X and Y.
- $0 \leq R^2 \leq 1$.
- R^2 is interpreted as the percentage of variability in Y captured by the regression model.

Reading R Output

```
> summary(model)

Call:
lm(formula = price ~ sqft, data = Sacramento)

Residuals:
    Min      1Q  Median      3Q     Max 
-231889  -54717  -11822   38993  600141 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13859.393   6948.714   1.995   0.0464 *  
sqft          138.546      3.796  36.495   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84130 on 930 degrees of freedom
Multiple R-squared:  0.5888,	Adjusted R-squared:  0.5884 
F-statistic:  1332 on 1 and 930 DF,  p-value: < 2.2e-16
```

R^2

PREDICTION

With the fitted model, we may want to predict $\hat{Y}^* = \mathbb{E}[Y|X^**]$ given a new value of X^*.

For the housing data, we predict the *average* price given the square footage of a new house by

$$\text{Price} = \hat{\beta}_0 + \hat{\beta}_1 \times \text{Square Footage}$$

We predict the *average* price of houses with a square footage of 2,000 by

$$13859.393 + 138.546 \times 2000 = \$290,951.4$$

Confidence Interval for Regression Line

A $(1-\alpha) \times 100\%$ confidence interval for $\hat{Y}^* = \mathbb{E}[Y|X^*]$ using the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ is

$$\hat{Y}^* \pm t^*_{\alpha/2,n-2}\hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(X^* - \bar{X})^2}{SXX}}$$

Sources of uncertainty:

- Regression parameters: $\hat{\beta}_0$ and $\hat{\beta}_1$ and
- Error variance $\hat{\sigma}^2$.

Prediction Intervals

The previous interval was a confidence interval for the average value $\hat{Y}^* = \mathbb{E}[Y|X^*]$ given a value of the predictor X^*.

- $\hat{Y}^* = \mathbb{E}[Y|X^*]$ is what we expect Y to be in long run when $X = X^*$.

Prediction Intervals

The previous interval was a confidence interval for the average value $\hat{Y}^* = \mathbb{E}[Y|X^*]$ given a value of the predictor X^*.

- $\hat{Y}^* = \mathbb{E}[Y|X^*]$ is what we expect Y to be in long run when $X = X^*$.
- $\hat{Y}^* = \mathbb{E}[Y|X^*]$ is *fixed* but unknown.

Prediction Intervals

The previous interval was a confidence interval for the average value $\hat{Y}^* = \mathbb{E}[Y|X^*]$ given a value of the predictor X^*.

- $\hat{Y}^* = \mathbb{E}[Y|X^*]$ is what we expect Y to be in long run when $X = X^*$.
- $\hat{Y}^* = \mathbb{E}[Y|X^*]$ is *fixed* but unknown.
- Values of Y when $X = X^*$ can vary around $\hat{Y}^*$.

Prediction Intervals

The previous interval was a confidence interval for the average value $\hat{Y}^* = \mathbb{E}[Y|X^*]$ given a value of the predictor X^*.

- $\hat{Y}^* = \mathbb{E}[Y|X^*]$ is what we expect Y to be in long run when $X = X^*$.
- $\hat{Y}^* = \mathbb{E}[Y|X^*]$ is *fixed* but unknown.
- Values of Y when $X = X^*$ can vary around $\hat{Y}^*$.
- Confidence intervals are for unknown parameters (e.g., $\hat{Y}^* = \mathbb{E}[Y|X^*]$).

Prediction Intervals

The previous interval was a confidence interval for the average value $\hat{Y}^* = \mathbb{E}[Y|X^*]$ given a value of the predictor X^*.

- $\hat{Y}^* = \mathbb{E}[Y|X^*]$ is what we expect Y to be in long run when $X = X^*$.
- $\hat{Y}^* = \mathbb{E}[Y|X^*]$ is *fixed* but unknown.
- Values of Y when $X = X^*$ can vary around $\hat{Y}^*$.
- Confidence intervals are for unknown parameters (e.g., $\hat{Y}^* = \mathbb{E}[Y|X^*]$).
- Prediction intervals are for random variables (e.g., Y^*) and will be wider than confidence intervals.

Prediction Intervals

A $(1-\alpha)\times 100\%$ prediction interval for the actual value Y^* when $X = X^*$ is

$$\hat{Y}^* \pm t^*_{\alpha/2,n-2}\hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(X^* - \bar{X})^2}{SXX}}$$

where $\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X^*$. Sources of uncertainty/variability:

- Regression parameters: $\hat{\beta}_0$ and $\hat{\beta}_1$ and
- Error variance $\hat{\sigma}^2$.
- Random fluctuation of actual values around the regression line.

Intervals in R

Use the `predict` function[1].

```
> model <- lm(price ~ sqft, data = Sacramento)
> predict(model, newdata = data.frame(sqft = 2000),
interval = "confidence", level = 0.95)
       fit        lwr      upr
1 290952.3 285043.1 296861.5
> predict(model, newdata = data.frame(sqft = 2000),
interval = "prediction", level = 0.95)
       fit        lwr      upr
1 290952.3 125748.3 456156.3
```

[1] See R notes on Canvas