Lecture 23

Part 5 Linear Regression

Simple Linear Regression

1. Checking the Model Assumptions

(c) Certain plots can be used for checking the independence of the residuals, but if the sample is collected properly (i.e., randomly) this hopefully shouldn't be a major problem.

- Can plot the residuals against the order in which the observations were collected to see if there is any time correlation between the residuals.

2. Testing Overall Significance of Model

- Once we are happy that the model assumptions are satisfied, the next thing we should do is test the *overall significance* of the model.
- Testing the overall significance of the model is equivalent to testing whether or not a model exists.
- That is, does a linear relationship between X and Y even exist.
- How might we be able to test this?

2. Testing Overall Significance of Model

- If $\beta_1 = 0$, what happens to the simple linear regression model?
- The model

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

  becomes

$$Y = \beta_0 + \epsilon$$

- That is, X disappears from the model, indicating that no linear relationship exists between X and Y.

Hypotheses

- The hypotheses for testing the overall significance of the simple linear regression model are:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- Note that we are testing the two-tailed alternative, since either $\beta_1 > 0$ or $\beta_1 < 0$ would indicate that a model exists.

Test Statistic

- The test statistic that we use to test the above hypotheses is the T-statistic:

$$T = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

where $s_{\hat{\beta}_1} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2}{(n-1)s_X^2}}$ is an estimate of the standard error of $\hat{\beta}_1$ (i.e., the standard deviation of the sampling distribution of $\hat{\beta}_1$).

Decision Rule

- For our decision rule, we need to compare this T-statistic to a t-distribution with $n-2$ degrees of freedom.
- Since it is a two-tailed test, we reject H_0 at a significance level of α if $T > t_{\frac{\alpha}{2},n-2}$ or $T < -t_{\frac{\alpha}{2},n-2}$.

Example

- Let's go back to our example to test the overall significance of the model which had attitude as the dependent variable Y and duration of residence as the independent variable X.
- Remember the hypotheses are:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

- We can calculate the test statistic by hand or we can also use the computer output.

Example

```
Call:
lm(formula = attitude ~ duration, data = city.dat)

Residuals:
Min      1Q  Median      3Q     Max
-1.9262 -0.7640 -0.4579  0.6165  2.7494

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.2237     1.0531     1.162 0.275114
duration     0.5585     0.0952     5.867 0.000239


Residual standard error: 1.493 on 9 degrees of freedom
Multiple R-squared: 0.7927,  Adjusted R-squared: 0.7697
F-statistic: 34.42 on 1 and 9 DF,  p-value: 0.0002387
```

Example

- The test statistic is calculated as:

$$T = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{0.5585}{0.0952} = 5.867$$

- There were $n = 11$ observations in our sample, so we compare this to a t-distribution with $n - 2 = 9$ degrees of freedom.
- At a significance level of $\alpha = 0.05$, the rejection region is therefore $T > 2.262$ or $T < -2.262$.

Example

- Since $5.867 > 2.262$, we reject H_0 and conclude that there is a significant linear relationship between X and Y.
- Alternatively, the computer output also gives us the p-value for the *two-tailed* alternative hypothesis.
- So we can reach the same conclusion by comparing the p-value of 0.000239 to $\alpha = 0.05$.

Testing the Correlation Coefficient

- Recall that the correlation coefficient ρ measured the strength of a linear relationship between two variables.
- We can also test the overall significance of the simple linear regression model by testing the following hypotheses:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Testing the Correlation Coefficient

- The test statistic we use is based on the sample correlation coefficient r:

$$T = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

- For the decision rule, we again compare this T-statistic to a t-distribution with $n-2$ degrees of freedom and reject H_0 at a significance level of α if $T > t_{\frac{\alpha}{2},n-2}$ or $T < -t_{\frac{\alpha}{2},n-2}$.

Testing the Correlation Coefficient

- Note that when testing $\beta_1 = 0$ in the simple linear regression model and when testing $\rho = 0$, both test statistics are compared to the same sampling distribution.
- These two tests are indeed equivalent, in the sense that some algebra can show:

$$\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

Testing the Correlation Coefficient

- For our city example data, the sample correlation coefficient between X and Y is $r_{XY} = 0.8903507$, so we get:

$$T = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

$$= \frac{0.8903507 \times \sqrt{11-2}}{\sqrt{1-0.8903507^2}}$$

$$= 5.867$$

General Test for β_1

- Aside from testing the overall significance of the simple linear regression model, we can also test more general hypotheses regarding β_1:

$$H_0 : \beta_1 = c$$

$$H_1 : \beta_1 (\neq, <, >) c$$

General Test for β_1

- The test statistic takes on the usual form:

$$T = \frac{\hat{\beta}_1 - c}{s_{\hat{\beta}_1}}$$

- For the decision rule, we compare the test statistic to a t-distribution with $n-2$ degrees of freedom.

General Test for β_0

- Although most interest in a simple linear regression usually concerns β_1, we can also test hypotheses regarding the intercept parameter β_0:

$$H_0 : \beta_0 = c$$

$$H_1 : \beta_0 (\neq, <, >) c$$

General Test for β_0

- The test statistic is:

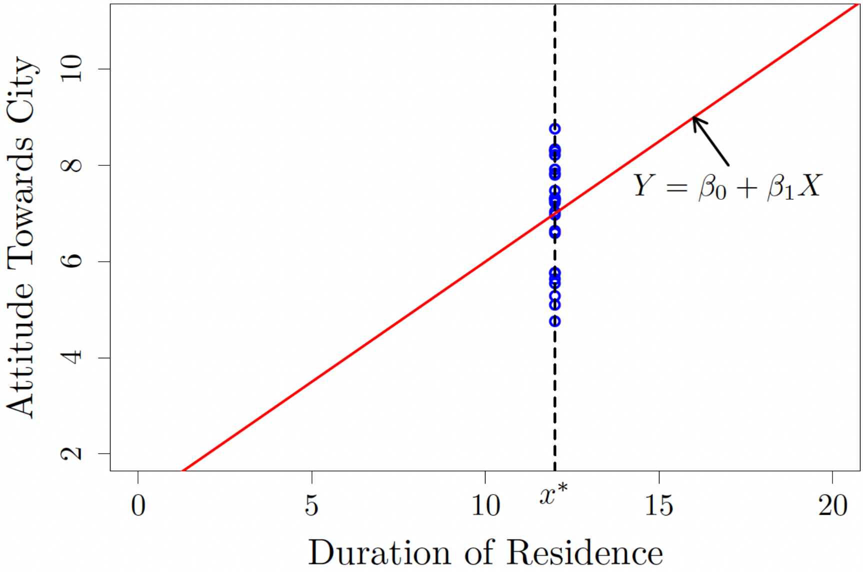
$$T = \frac{\hat{\beta}_0 - c}{s_{\hat{\beta}_0}}$$

  where $s_{\hat{\beta}_0} = s_{\hat{\beta}_1} \times \sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n}}$ is an estimate of the standard error of $\hat{\beta}_0$.
- For the decision rule, we again compare the test statistic to a t-distribution with $n-2$ degrees of freedom.

3. Estimating σ_ϵ^2

- Now that we have established that the overall model is significant, how good is our model?
- Recall that the error variable ϵ_i represents the difference between the Y_i value of each observation and the straight line component of the regression model.
- Further, the model assumptions stated that $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$.

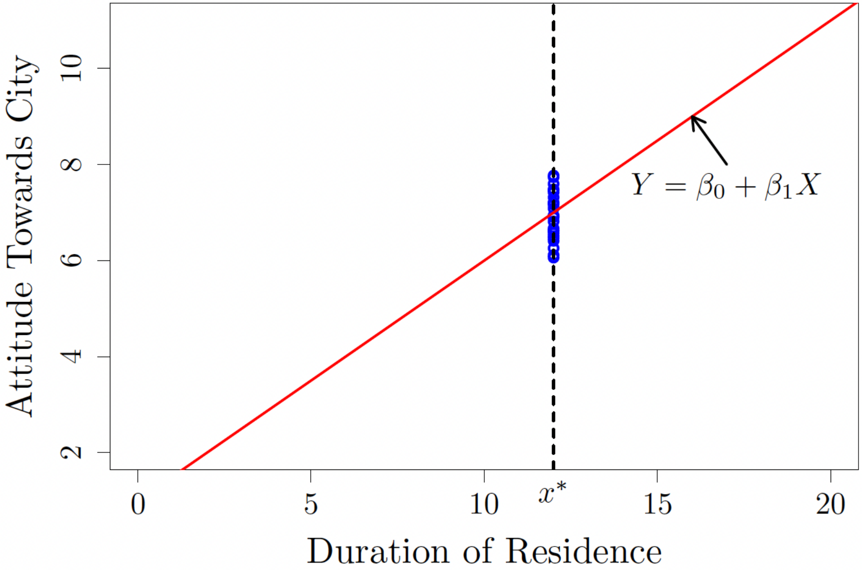
3. Estimating σ_ϵ^2



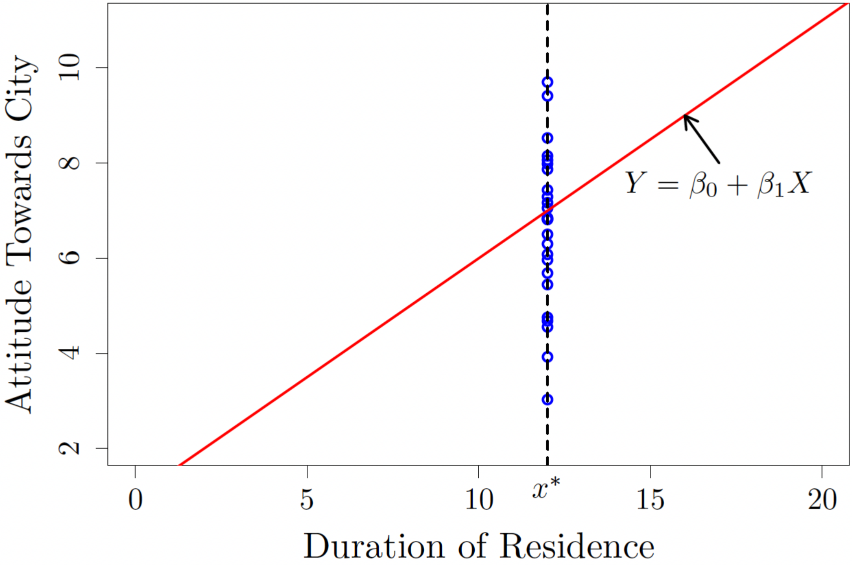
3. Estimating σ_ϵ^2

- If σ_ϵ^2 is small, the errors ϵ_i are close to the mean 0, indicating that the regression model fits the data well.
- If σ_ϵ^2 is large, some of the errors ϵ_i will be large, indicating that the regression model does not fit the data well.
- But σ_ϵ^2 is an unknown population parameter, which therefore has to be estimated.

3. Estimating σ_ϵ^2



3. Estimating σ_ϵ^2



3. Estimating σ_ϵ^2

- Again, since we don't know the true errors ϵ_i, we use the residuals e_i to obtain an *unbiased* estimator of σ_ϵ^2:

$$s_\epsilon^2 = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{\sum_{i=1}^n \left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

- Note that s_ϵ, the square root of s_ϵ^2, is called the **standard error of estimate**.

3. Estimating σ_ϵ^2

- Just based on the value of s_ϵ^2, it can be difficult to determine whether it's small enough to indicate a good model.
- However, it is useful for comparing two different models - the model with the smaller standard error of estimate is generally considered better.

4. Calculating R^2

- If the overall model is significant, there exists a linear relationship between X and Y.
- Would be nice to be able to measure the *strength* of the linear relationship.
- This is measured by R^2, the **coefficient of determination**, and is defined by:

$$R^2 = \frac{s_{XY}^2}{s_X^2 s_Y^2}$$

4. Calculating R^2

- Note that R^2 is just the square of the sample correlation coefficient.
- There is also another way to express R^2, which is based on how much variation is explained by the regression model.
- Just like ANOVA, we can define some sums of squares...

Sums of Squares

- Total sum of squares:

$$SS(Total) = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Sum of squares for regression:

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

- Sum of squares for error:

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

4. Calculating R^2

- It is also true that:

$$SS(Total) = SSR + SSE$$

- Some algebra will show that the coefficient of determination can also be written as:

$$R^2 = \frac{SSR}{SS(Total)}$$

4. Calculating R^2

- Therefore, R^2 also measures the proportion of total variation in Y that is explained by the simple linear regression model.
- Similar to s_ϵ^2, R^2 is useful for comparing different models.

Using the Model

- Once we have assessed the simple linear regression model and concluded that it was appropriate for our data, we can proceed to use our estimated model.
- Suppose we have a new observation from the population with an X value equal to $X = x_g$.
- Given this value x_g, we can *predict* the value of Y using our estimated model:

$$\hat{y}_g = \hat{\beta}_0 + \hat{\beta}_1 x_g$$

Point Estimate

- So $\hat{y}_g$ gives us a *point estimate* for the value of Y when $X = x_g$.
- However, it does not tell us anything about how close this predicted value is to the true value of Y.
- How can we address this problem?
- We can use an interval estimator!
- Before we derive some interval estimators, for a given value of $X = x_g$, there are actually two different quantities that we might be interested in estimating...

Point Estimate

1. The particular value of Y for that particular observation with $X = x_g$, that is,
$$y_g = \beta_0 + \beta_1 x_g + \epsilon_g$$
2. The *expected value* of Y for *all* observations with $X = x_g$, that is,
$$E(Y|X = x_g) = \beta_0 + \beta_1 x_g$$

- For both these quantities, we use $\hat{y}_g$ as our point estimator, but we use slightly different interval estimators.

Confidence Intervals

- For a given value of $X = x_g$,

1. The confidence interval for a particular value of Y (also called the *prediction interval*) is given by:

$$\hat{y}_g \pm t_{\frac{\alpha}{2},n-2} \times s_\epsilon \sqrt{1 + \frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_X^2}}$$

2. The confidence interval for the expected value of Y is given by:

$$\hat{y}_g \pm t_{\frac{\alpha}{2},n-2} \times s_\epsilon \sqrt{\frac{1}{n} + \frac{(x_g - \bar{X})^2}{(n-1)s_X^2}}$$

Confidence Intervals

- The intervals look very similar, the only change being the term within the square root.
- For the same confidence level, the confidence interval for a particular value of Y (prediction interval) is wider than the confidence interval for the expected value of Y.
- This is because there is more variability associated with predicting a particular value of Y than there is with estimating a mean or expected value.