Ch 2 - Simple Linear Regression

Matthew Wascher

Simple Linear Regression Model

Suppose we have a single response variable Y and a single predictor variable X. The **simple linear regression model** characterizes the relationship between X and Y by

$$E(Y|X) = \beta_0 + \beta_1 X$$
$$Var(Y|X) = \sigma^2$$

Simple Linear Regression Model

Suppose we have observed data $(x_1, y_1) \ldots (x_n, y_n)$. Another way of stating this model is

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

where the e_i are iid random variables with $E(e_i) = 0$ and $Var(e_i) = \sigma^2$.

Simple Linear Regression Model

We can see these are equivalent by computing

$$E(Y|X = x_i) = \beta_0 + \beta_1 x_i$$

$$Var(Y|X = x_i) = \sigma^2$$

The Error Term

The error term e_i is the true value of y_i minus the expected value of y_i

$$e_i = y_i - E(Y|X = x_i) = y_i - (\beta_0 + \beta_1 x_i)$$

We make two further assumptions about the e_i

1. $E(e_i|X = x_i) = 0$. The mean of e_i is 0 for every possible x_i.
2. The e_i form an independent collection.

Common stronger assumptions include that the e_i are independent of the x_i (replacing 1.) and that the e_i are Normally distributed. Stronger assumptions are needed for things like tests and confidence intervals, but not to derive the basic regression model.

Estimation

The goal of simple linear regression is to estimate the model parameters β_0, β_1, and σ^2. Estimators are denoted with hats, in this case $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma^2}$

We will choose $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\sigma^2}$ to best “fit” the observed data. There are many ways to measure “fit”.

Fitted values and residuals

For estimated parameters $\hat{\beta}_0$ and $\hat{\beta}_1$, we define

1. The fitted values $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.
2. The residuals $\hat{e}_i = y_i - \hat{y}_i$.
   These are the plug-in estimators for y_i and e_i using the estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$.

Least Squares Estimation

One way to choose $\hat{\beta}_0$ and $\hat{\beta}_1$ is with the **least squares criterion** leading to **Ordinary Least Squares (OLS)** regression. The least squares criterion says choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to be the values of β_0 and β_1 that minimize

$$RSS = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i) \Big)^2$$

RSS stands for residual sum of squares and represents the sum of the squared vertical distances from y_i to the fitted value $\hat{y}_i$.

Deriving OLS Regression

We can derive the least squares estimates of β_0 and β_1 using multivariate calculus. First, we find the critical point where

$$
\begin{aligned}
\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i) \Big)^2 &= 0 \\
\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 x_i) \Big)^2 &= 0
\end{aligned}
\tag{1}
$$

We then verify that this is the minimum by checking that $|\mathbf{H}| > 0$, where $\mathbf{H}$ is the Hessian matrix. See Appendix A.3 of the book for the full derivation.

Formulas for OLS Regression

The following are formulas for the least squares estimates of β_0 and β_1

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned} \tag{2}
$$

While we will use technology to compute these estimates when working with data, these formulas are useful for understanding the properties of the estimators.

Fitting OLS Regression in R

We can easily compute $\hat{\beta}_0$ and $\hat{\beta}_1$ in R using the function lm(). We can call

```
fit = lm(y ~ x)
summary(fit)
```

This reports $\hat{\beta}_0$ (the intercept) and $\hat{\beta}_1$ (the slope associated with the predictor variable x).

Estimating σ^2

We estimate σ^2 by

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

The $n-2$ is because we lose 2 degrees of freedom from estimating $\hat{\beta}_0$ and $\hat{\beta}_1$. $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ is called the **residual standard error.**

Properties of OLS Estimators

1. Under our regression assumptions $E(\hat{\beta}_1|X) = \beta_1$.

$$
\begin{aligned}
E(\hat{\beta}_1|X) &= E\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \\
&= \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2} E\left(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right) \\
&= \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2} E\left(\sum_{i=1}^{n}(x_i-\bar{x})(\beta_0+\beta_1 x_i+e_i-(\beta_0+\beta_1\bar{x}+\bar{e}))\right) \\
&= \frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2} E\left(\sum_{i=1}^{n}(x_i-\bar{x})(\beta_1(x_i-\bar{x})+e_i-\bar{e})\right) \\
&= \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} E(\beta_1) + E\left(\sum_{i=1}^{n}(x_i-\bar{x})(e_i-\bar{e})\right) \\
&= \beta_1 + 0 = \beta_1
\end{aligned}
$$

Properties of OLS Estimators

2. Under our regression assumptions $E(\hat{\beta}_0|X) = \beta_0$.

$$\begin{aligned} E(\hat{\beta}_0|X) &= E(\bar{y} - \hat{\beta}_1\bar{x}) \\ &= \bar{y} - \bar{x}E(\hat{\beta}_1) \\ &= \bar{y} - \bar{x}\beta_1 \text{ (since } \hat{\beta}_1 \text{ is unbiased)} \\ &= \beta_0 \end{aligned}$$

Properties of OLS Estimators

3. Under our regression assumptions $E(\hat{\sigma}^2|X) = \hat{\sigma}^2$.

The proof is beyond the scope of our class and involves χ^2 distributions.

Properties of OLS Estimators

4. Under our regression assumptions

- $Var(\hat{\beta}_1|X) = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$.
- $Var(\hat{\beta}_0|X) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$.
- $Cov(\hat{\beta}_1, \hat{\beta}_0|X) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Standard Errors

We may wish to estimate the variances or standard errors $\hat{\beta}_0$ and $\hat{\beta}_1$. We substitute $\hat{\sigma}^2$ for σ and obtain the plug-in estimators

- $\hat{Var}(\hat{\beta}_1|X) = \hat{\sigma}^2 \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$.
- $se(\hat{\beta}_1|X) = \sqrt{\hat{Var}(\hat{\beta}_1|X)}$.
- $\hat{Var}(\hat{\beta}_0|X) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$.
- $se(\hat{\beta}_0|X) = \sqrt{\hat{Var}(\hat{\beta}_0|X)}$.

These standard errors are output as part of summary(fit) in R.

Normal Errors

If we assume the e_i are iid Normal random variables with mean 0 and variance σ^2, we can develop tests and confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$.

In this case $\hat{\beta}_0$ and $\hat{\beta}_1$ are also Normally distributed. (Why?)

t Tests and Confidence Intervals

Recall that if a random variable U follows a Normal distribution then

$$\frac{U - \mu_U}{\hat{\sigma}_U} \sim t_{df}$$

where *df* is the degrees of freedom of $\hat{\sigma}_U$.

t Tests and Confidence Intervals

Using the fact that $\hat{\sigma}^2$ has $n-2$ degrees of freedom we can conclude

$$\begin{aligned} \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1|X)} &\sim t_{n-2} \\ \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0|X)} &\sim t_{n-2} \end{aligned} \tag{3}$$

t Tests and Confidence Intervals

Thus to test the hypotheses

$$H_0 : \beta_1 = \beta_1^*$$

$$H_1 : \beta_1 \neq \beta_1^*$$

we obtain the test statistic $t^* = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1|X)}$ and the p-value $P(t_{n-2} > |t^*|)$.

To test the hypotheses

$$H_0 : \beta_0 = \beta_0^*$$

$$H_1 : \beta_0 \neq \beta_0^*$$

we obtain the test statistic $t^* = \frac{\hat{\beta}_0 - \beta_0}{se(\hat{\beta}_0|X)}$ and the p-value $P(t_{n-2} > |t^*|)$.

t Tests and Confidence Intervals

summary(fit) in R reports $|t^*|$ and the p-value $P(t_{n-2} > |t^*|)$ for the tests

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

and

$$H_0 : \beta_0 = 0$$

$$H_1 : \beta_0 \neq 0$$

In particular, the p-value from the first test is often used as a measure of evidence that X is linearly associated with (and thus useful for predicting) Y.

t Tests and Confidence Intervals

$100(1-\alpha)\%$ confidence intervals can be obtained by

$$\hat{\beta}_1 \pm t_{1-\alpha/2,n-2} se(\hat{\beta}_1)$$

and

$$\hat{\beta}_0 \pm t_{1-\alpha/2,n-2} se(\hat{\beta}_0)$$

Prediction

An important use of regression models is predicting the value of the response y^* for a new value of the predictor x^*. We can observe

$$y^* = \beta_0 + \beta_1 x^* + e^*$$

Thus $\tilde{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x^*$ is an unbiased estimator of y^* since

$$E(\tilde{y}^* | x^*) = \beta_0 + \beta_1 x^* = y^*$$

Prediction

We also wish to quantify the uncertainty associated with our prediction.

$$Var(\tilde{y}^*|x^*) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)$$

Prediction Intervals

If the e_i are Normaly distributed then so is $\tilde{y}^*$ (Why?). A confidence interval for y^* is called **prediction interval**. The $100(1-\alpha)$ prediction interval for y^* given x^* is

$$\tilde{y}^* \pm t_{1-\alpha/2,n-2} se(\tilde{y}^*|x^*)$$

Prediction in R

Prediction in R can be performed for a wide variety of models using the powerful predict() function. Suppose we have fit a model (for example using fit = lm(y x̃).) Let newx be a data frame containing the x^* where we want predictions of y^*. We can obtain these by

predict(fit, newx)

We can obtain prediction intervals using

predict(fit, newdata, interval = ”prediction”)

Confidence Intervals for Fitted Values

We can also express our uncertainty about our estimate $\hat{y}$ of $E(Y|x)$. We can observe that

$$Var(\hat{y}|x) = \sigma^2 \left(\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right)$$

Confidence Intervals for Fitted Values

We could obtain a confidence interval for a particular $x = x_i$ using a t interval, but it is more common to create a confidence band for all x value simultaneously. The $100(1-\alpha)\%$ confidence band is

$$(\hat{\beta}_0 + \hat{\beta}_1 x) \pm \left(2F_{\alpha,2,n-2}\right)^2 se(\hat{y}|x)$$

$F_{2,n-2}$ is an F distribution with 2 and $n-2$ degrees of freedom.

Confidence Bands in R

We can obtain confidence bands in by defining a data frame grid containing a fine grid of x values and using

predict(fit, grid, interval = ”confidence”)

The Coefficient of Determination

The OLS line is the "best" fit in some sense, but how good is it? One measure is the **coefficient of determination** R^2. It is defined by

$$R^2 = 1 - \frac{\sum_{i=1}^n \hat{e}_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

R^2 measures the proportion of variation in y explained by the model using y. $\sum_{i=1}^n (y_i - \bar{y})^2$ is the total variation in y while $\sum_{i=1}^n \hat{e}_i^2$ remaining variation after fitting the OLS model using x. This is the "Multiple R-Squared" reported in summary(fit) in R.

Correlation

The **correlation between** x **and** y, written r_{xy} is a measure of the strength and direction of the *linear* relationship between x and y. It is related to R^2 by

$$r_{xy} = \sqrt{R^2}$$

with the sign of the square root being determined by the sign of $\hat{\beta}_1$.

Residuals

The residuals (the $\hat{e}_i$) are useful for checking our assumptions about the e_i, such as whether the e_i are mean 0, constant variance, or Normally distributed. There are many ways one can use the residuals, some of which we will touch on throughout this class. One of the most basic is plotting the residuals vs the x or the fitted values.