Linear regression in R

Topic 4.4. Prediction intervals; The 'why' & 'who cares' of regression

Sonja Petrović

Created for ITMD/ITMS/STAT 514

Spring 2021.

Goals for this lecture

- Understand & interpret intervals (confidence & prediction) in multiple and simple linear regression
- Fit a regression model in `Python`
- Understand & interpret `Python` output for linear models
- Many other considerations for regression modeling
  - qualitative predictors [another semester]
  - extensions of the linear model [removing additive assumption; non-linear relationships]
  - potential problems [simple overview].

Prediction using regression

The starting point of regression

What are we *really* modeling here?

- Previously:
  - estimating parameters based on an iid sample $Y_1, \ldots, Y_n$
  - $\implies E[Y_1] = \cdots = E[Y_n]$.
  - in particular, $E[Y]$ does not depend on athe value of any other variable.
- Regression:

The starting point of regression

What are we *really* modeling here?

- Previously:
  - estimating parameters based on an iid sample $Y_1, \ldots, Y_n$
  - $\implies E[Y_1] = \cdots = E[Y_n]$.
  - in particular, $E[Y]$ does not depend on athe value of any other variable.
- Regression:
  - random variable Y has a mean that depends on (one or several) non-random vars $X_1, \ldots, X_p$ (predictors)
  - Deterministic model: $Y = \beta_0 + \beta_1 X$
  - **Probabilistic model**: $E[Y] = \beta_0 + \beta_1 X$. Equivalently:

The starting point of regression

What are we *really* modeling here?

- Previously:
  - estimating parameters based on an iid sample $Y_1, \ldots, Y_n$
  - $\implies E[Y_1] = \cdots = E[Y_n]$.
  - in particular, $E[Y]$ does not depend on athe value of any other variable.
- Regression:
  - random variable Y has a mean that depends on (one or several) non-random vars $X_1, \ldots, X_p$ (predictors)
  - Deterministic model: $Y = \beta_0 + \beta_1 X$
  - **Probabilistic model**: $E[Y] = \beta_0 + \beta_1 X$. Equivalently:

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\substack{\text{deterministic} \\ \text{output}}} + \underbrace{\epsilon}_{\substack{\text{random} \\ \text{output}}}$$

  - Least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ minimize the RSS.

$\rightarrow$ so... now what??

Recall this example from previous lecture

- `Auto` data set, regression on `Y=mpg` vs. `X=horsepower`.

```
fit.lm <- lm(mpg ~ horsepower, data=Auto)
```

- What is the predicted `mpg` associated with a `horsepower` of 98? What are the associated 95% confidence and prediction intervals?

```
new <- data.frame(horsepower = 98)
predict(fit.lm, new) # predicted mpg
```

```
       1
24.46708
```

```
predict(fit.lm, new, interval="confidence") # conf interval
```

```
       fit      lwr      upr
1 24.46708 23.97308 24.96108
```

```
predict(fit.lm, new, interval="prediction") # pred interval
```

```
       fit     lwr      upr
1 24.46708 14.8094 34.12476
```

- → confidence interval vs. prediction interval ←

Confidence vs. prediction intervals

Three sorts of uncertainty associated with the prediction of Y based on $X_1, \ldots, X_p$:

- $\hat{\beta}_i \approx \beta_i$: least squares plane is **an estimate** for the true regression plane.
  - reducible error
- assuming a linear model for $f(X)$ is usually an approximation of reality
  - model bias [potential reducible error?]
  - to operate here, we ignore this discrepancy
- even if we knew true β_i, still no perfect knowledge of Y because of random error ϵ
  - irreducible error
  - how much will Y vary from $\hat{Y}$?
  - we use *prediction intervals*. Always wider than confidence intervals.

Example: `Advertising` confidence

Confidence interval

Quantify the uncertainty surrounding the average sales over a large number of cities.

For example:

- given that \$100,000 is spent on TV advertising and
- \$20,000 is spent on radio advertising in each city,
- the 95 % confidence interval is [10,985, 11,528].
- We interpret this to mean that 95 % of intervals of this form will contain the true value of $f(X)$.

Example: `Advertising` prediction

Prediction interval

Can be used to quantify the uncertainty surrounding sales for a particular city.

- Given that \$100,000 is spent on TV advertising and
- \$20,000 is spent on radio advertising in that city
- the 95 % prediction interval is [7,930, 14,580].
- We interpret this to mean that 95 % of intervals of this form will contain the true value of Y for this city.

$\rightarrow$ Note that both intervals are centered at 11,256, but that the prediction interval is substantially wider than the confidence interval, reflecting the increased uncertainty about sales for a given city in comparison to the average sales over many locations.

Other considerations and extensions

Qualitative predictors

This is out of scope this semester (we are out of time!), but consider this setup:

- there may be a *qualitative predictor* (that, is a discrete random variable X_i) – it's also called a **factor**
- suppose X_i has only two levels (e.g. female and not female)
- we use a *dummy variable*

$$x_i = \begin{cases} 1, \text{ if } i\text{th person is female} \\ 0, \text{ if } i\text{th person is not female} \end{cases}$$

- use this as predictor in the regression equation.

The model becomes:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, \text{ if } i\text{th person is female} \\ \beta_0 + \epsilon_i, \text{ if } i\text{th person is not female} \end{cases}$$

The model becomes:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon = \begin{cases} \beta_0 + \beta_1 + \epsilon_i, \text{ if } i\text{th person is female} \\ \beta_0 + \epsilon_i, \text{ if } i\text{th person is not female} \end{cases}$$

Interpret:

- β_0 = average Y among non-females
- $\beta_0 + \beta_1$ = average Y among females
- β_1 average difference in Y between the two groups.

Extensions of the linear model

What is wrong with the linear model? It works quite well!

Yes – but sometimes the (restrictive) assumptions are violated in practice.

Assumption 1: additivity

The relationship between the predictors and response is additive.

- effect of changes in a predictor X_j on the response Y is independent of the values of the other predictors.

Assumption 2: linearity

The relationship between the predictors and response is linear.

- the change in the response Y due to a one-unit change in X_j is constant, regardless of the value of X_j.

Removing the additive assumption

Previous analysis of `Advertising` data: both `TV` and `radio` seem associated with sales.

- The linear models that formed the basis for this conclusion:

$$sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \epsilon$$

- We will now [in the notes] explain how to augment this model by allowing interaction between `radio` and `TV` in predicting `sales`:

Removing the additive assumption

Previous analysis of `Advertising` data: both `TV` and `radio` seem associated with sales.

- The linear models that formed the basis for this conclusion:

$$sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \epsilon$$

- We will now [in the notes] explain how to augment this model by allowing interaction between `radio` and `TV` in predicting `sales`:

$$sales = \beta_0 + \beta_1 \times TV + \beta_2 \times radio + \beta_3 \times (radio \times TV) + \epsilon$$
$$= \beta_0 + (\beta_1 + \beta_3 \times radio) \times TV + \beta_2 \times radio + \epsilon. \quad (3.33)$$

Interpretation:

- β_3 = increase in the effectiveness of TV advertising for a one unit increase in radio advertising (or vice-versa).

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

TABLE 3.9. *For the* Advertising *data, least squares coefficient estimates associated with the regression of* sales *onto* TV *and* radio*, with an interaction term, as in (3.33).*

Figure 1: ISLR Table 3.9

Discuss:

- main effects
- hierarchical principle

Removing the linear assumption

Polynomial regression

models non-linear relationships.

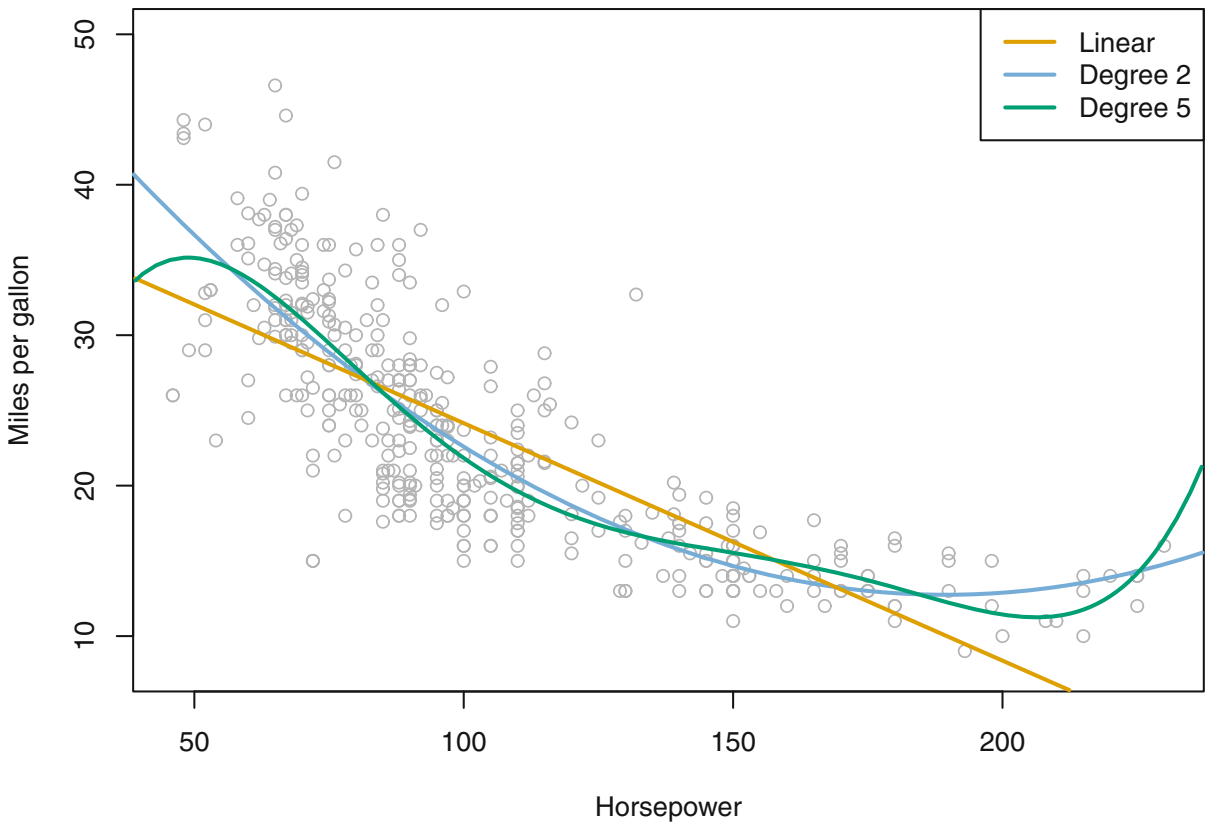

Figure 2: ISLR Fig. 3.8. The `Auto` data set. For a number of cars, `mpg` and `horsepower` are shown. The linear regression fit is shown in orange. The linear regression fit for a model that includes `horsepower^2` is shown as a blue curve. The linear regression fit for a model that includes all polynomials of `horsepower` up to fifth-degree is shown in green.

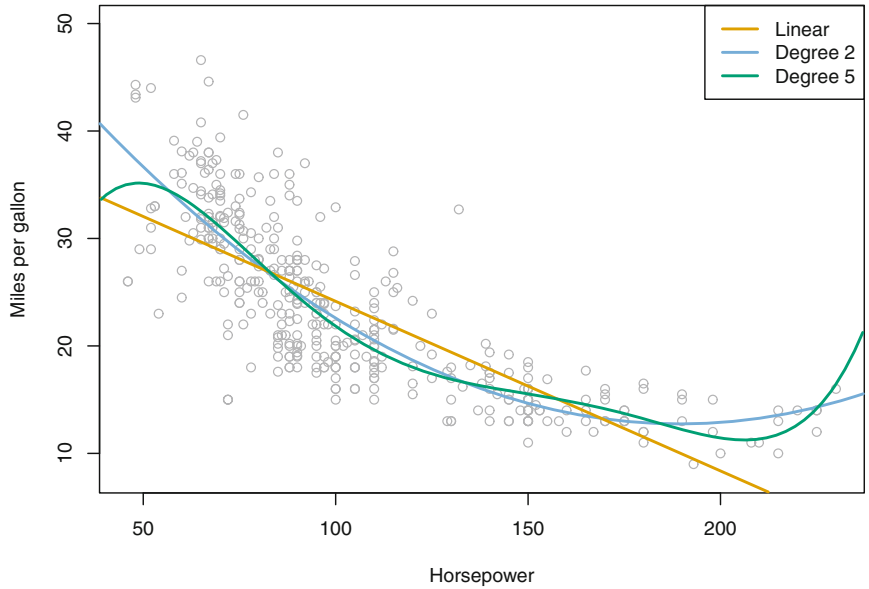

Figure 3: ISLR Fig. 3.8.

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	56.9001	1.8004	31.6	< 0.0001
`horsepower`	-0.4662	0.0311	-15.0	< 0.0001
`horsepower`^{2}	0.0012	0.0001	10.1	< 0.0001

TABLE 3.10. *For the* `Auto` *data set, least squares coefficient estimates associated with the regression of* `mpg` *onto* `horsepower` *and* `horsepower`^{2}.

Figure 4: ISLR Table 3.10

Potential problems

Common issues and problems

1. Non-linearity of the response-predictor relationships. 2. Correlation of error terms.
2. Non-constant variance of error terms.
3. Outliers.
4. High-leverage points.
5. Collinearity

(You will learn to deal with these in another course that focuses more on using regression in your application domain.)

License

Contents of this lecture is based on the chapter 3 of the textbook Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *' An Introduction to Statistical Learning: with Applications in R'*.