6.4 Prediction and Residual Analysis

6.4.1 Prediction Error and Interval

1. Consider the estimated model:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

2. For $c_1, c_2, \cdots, c_k$, a data point, the predicted value of $\hat{y}$ is

$$\hat{\theta} = E(y | x_1 = c_1, x_2 = c_2, \ldots, x_k = c_k)$$

   To get a confidence interval on this estimate, we need a way of calculating its standard error.

(a) Run this model instead:

$$y = \theta_0 + \beta_1(x_1 - c_1) + \beta_2(x_2 - c_2) + \cdots + \beta_k(x_k - c_k) + u$$

(b) Estimate of $\hat{\theta}$ will be $\hat{\theta}_0$ and its CI can be obtained using its standard error.

(c) Note: when $c_1 = \bar{x}_1$ and so on, you get the smallest SE and θ_0 merely measures the part of variation in dependent variable in your sample which you could explain.

3. Now consider a value in data y^0 *but not in our sample*. In order to obtain a prediction error (out of sample error) we calculate what our regression predicts its value would be and what its value actually is:

$$\hat{e}^0 = y^0 - \hat{y}^0$$

4. Prediction error will have a standard deviation, which gives us the CI for our perdiction. The **variance of our prediction error** is :

$$Var(\hat{e}^0) = Var(\hat{y}^0) + Var(u^0)$$

where $Var(\hat{y}^0)$ is the sample variance in your data and depends upon how large your sample is.

5. $Var(u^0)$ does not depend on your sample and is unobservable σ^2 However, we can use sample estimate of unobservable σ^2 to obtain

$$se(\hat{e}_0) = \left\{ [se(\hat{y}^0)]^2 + \hat{\sigma}^2 \right\}^{\frac{1}{2}}$$

$se(\hat{e}_0)$ has t-distribution with $n - k - 1$ degrees of freedom. This gives us the prediction interval as:

$$\hat{y}^0 \pm t_{0.025} \cdot se(\hat{e}_0)$$

where $t_{0.025} \approx 1.96$ and the above CI gives a 95% interval.

6. Example: Consider a model which tries to explain *colgpa* which is the cumulative grade point average, *hsize* is the size of the class graduating, *hsperc* is academic percentile in graduating class, *sat* is combined SAT scores, *female* is gender binary and *athlete* is a participation binary from GPA2:

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u$$

   (a) Obtain a 95% confidence interval for the expected college GPA when sat = 1200, hsperc = 30 and hsize = 5 (try another combination too!)

   (b) Also obtain a 95% confidence interval for any value of sat, hsperc and hsize

```
> library(wooldridge)
> data(gpa2)
> # Regression equation 6.32
> summary(lm(colgpa ~ sat + hsperc + hsize + hsizesq))
Error in eval(predvars, data, env) : object 'colgpa' not found
> summary(lm(colgpa ~ sat + hsperc + hsize + hsizesq, gpa2))

Call:
lm(formula = colgpa ~ sat + hsperc + hsize + hsizesq, data = gpa2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.57543 -0.35081  0.03342  0.39945  1.81683

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.493e+00  7.534e-02  19.812  < 2e-16 ***
sat          1.492e-03  6.521e-05  22.886  < 2e-16 ***
hsperc      -1.386e-02  5.610e-04 -24.698  < 2e-16 ***
hsize       -6.088e-02  1.650e-02  -3.690 0.000228 ***
hsizesq      5.460e-03  2.270e-03   2.406 0.016191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5599 on 4132 degrees of freedom
Multiple R-squared:  0.2781,    Adjusted R-squared:  0.2774
F-statistic:   398 on 4 and 4132 DF,  p-value: < 2.2e-16
```

```
> # To obtian confidence interval of our predictions on colgpa conditional on our x's we need to run
 another regression.
> # JW runs them at arbitrary values of x's, but let us run them at sample average deviations
> summary(gpa2)
      sat            tothrs          colgpa          athlete          verbmath
 Min.   : 470   Min.   :  6.00   Min.   :0.000   Min.   :0.00000   Min.   :0.2597
 1st Qu.: 940   1st Qu.: 17.00   1st Qu.:2.210   1st Qu.:0.00000   1st Qu.:0.7759
 Median :1030   Median : 47.00   Median :2.660   Median :0.00000   Median :0.8667
 Mean   :1030   Mean   : 52.83   Mean   :2.653   Mean   :0.04689   Mean   :0.8805
 3rd Qu.:1120   3rd Qu.: 80.00   3rd Qu.:3.120   3rd Qu.:0.00000   3rd Qu.:0.9649
 Max.   :1540   Max.   :137.00   Max.   :4.000   Max.   :1.00000   Max.   :1.6667
     hsize           hsrank           hsperc           female           white
 Min.   :0.03   Min.   :  1.00   Min.   : 0.1667   Min.   :0.0000   Min.   :0.0000
 1st Qu.:1.65   1st Qu.: 11.00   1st Qu.: 6.4328   1st Qu.:0.0000   1st Qu.:1.0000
 Median :2.51   Median : 30.00   Median :14.5833   Median :0.0000   Median :1.0000
 Mean   :2.80   Mean   : 52.83   Mean   :19.2371   Mean   :0.4496   Mean   :0.9255
 3rd Qu.:3.68   3rd Qu.: 70.00   3rd Qu.:27.7108   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :9.40   Max.   :634.00   Max.   :92.0000   Max.   :1.0000   Max.   :1.0000
     black            hsizesq
 Min.   :0.00000   Min.   : 0.0009
 1st Qu.:0.00000   1st Qu.: 2.7225
 Median :0.00000   Median : 6.3001
 Mean   :0.05535   Mean   :10.8535
 3rd Qu.:0.00000   3rd Qu.:13.5424
 Max.   :1.00000   Max.   :88.3600
```

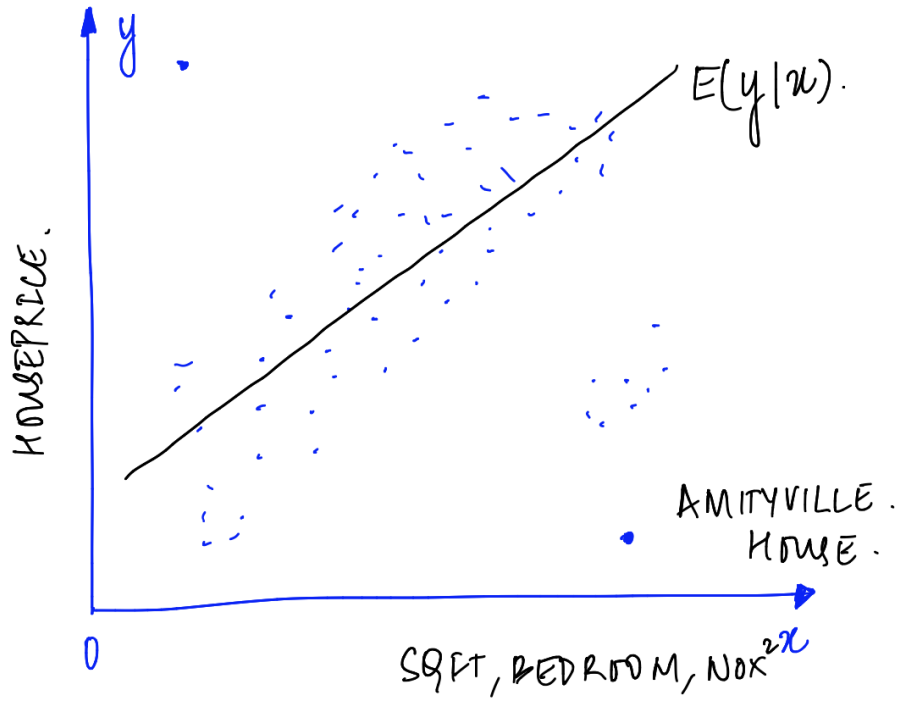
```
> summary(lm(colgpa ~ I(sat - 1030) + I (hsperc - 19.2371) + I(hsize - 2.80) + I(hsizesq - 10.8535),
 gpa2))

Call:
lm(formula = colgpa ~ I(sat - 1030) + I(hsperc - 19.2371) + I(hsize -
    2.8) + I(hsizesq - 10.8535), data = gpa2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.57543 -0.35081  0.03342  0.39945  1.81683

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.652e+00  8.704e-03 304.692  < 2e-16 ***
I(sat - 1030)         1.492e-03  6.521e-05  22.886  < 2e-16 ***
I(hsperc - 19.2371)  -1.386e-02  5.610e-04 -24.698  < 2e-16 ***
I(hsize - 2.8)       -6.088e-02  1.650e-02  -3.690 0.000228 ***
I(hsizesq - 10.8535)  5.460e-03  2.270e-03   2.406 0.016191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5599 on 4132 degrees of freedom
Multiple R-squared:  0.2781,    Adjusted R-squared:  0.2774
F-statistic:   398 on 4 and 4132 DF,  p-value: < 2.2e-16

> # The desired standard error is the SE for the intercept i.e. 8.704e-03
> A 95% CI is assuming normality 2.65 +- (1.96) * (8.704e-03)
Error: unexpected numeric constant in "A 95"
> # A 95% CI is assuming normality 2.65 +- (1.96) * (8.704e-03)
```

```
> # we can repeate the exercise of book as:
> summary(lm(colgpa ~ I(sat - 1200) + I (hsperc - 30) + I(hsize - 5) + I(hsizesq - 25), gpa2))

Call:
lm(formula = colgpa ~ I(sat - 1200) + I(hsperc - 30) + I(hsize -
    5) + I(hsizesq - 25), data = gpa2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.57543 -0.35081  0.03342  0.39945  1.81683

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.700e+00  1.988e-02 135.833  < 2e-16 ***
I(sat - 1200)    1.492e-03  6.521e-05  22.886  < 2e-16 ***
I(hsperc - 30)  -1.386e-02  5.610e-04 -24.698  < 2e-16 ***
I(hsize - 5)    -6.088e-02  1.650e-02  -3.690 0.000228 ***
I(hsizesq - 25)  5.460e-03  2.270e-03   2.406 0.016191 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5599 on 4132 degrees of freedom
Multiple R-squared:  0.2781,    Adjusted R-squared:  0.2774
F-statistic:   398 on 4 and 4132 DF,  p-value: < 2.2e-16

> # we get the exact same estimates as the ones in text.
```

```
> mod =lm(colgpa ~ I(sat - 1200) + I (hsperc - 30) + I(hsize - 5) + I(hsizesq - 25), gpa2)
```

6.4.2 Residual Analysis

1. **Residual analysis**: is when you look at the residual from a regression line to understand observations which are not fit well by the regression i.e. observations away from the line of fit.



What if the outlier is the Amityville House? Does that make sense?

You can live in the 'Amityville Horror' home for $850,000

Published: July 2, 2016 at 10:55 a.m. ET

By Daniel Goldstein

112

Allegedly haunted Long Island house was the scene of grisly mass murder in 1974

2. Example: Using the data in HPRICE1, we run a regression of *price* on *lotsize*, $sqrft$, and *bdrms* we uncovered the ghost house

   - **n**: sample of 88 homes
   - **Residuals**: the most negative residual is 2120.206, for the 81st house
   - **price**: the asking price for this house is \$120,206 below its predicted price

6.4.3 Predicting y when using log(y)

1. **Log transformations**: typically bring large values closer to down and inflate very small values. Thus it changes the standard deviation σ of the variable which is being transformed.

2. Consider the semi-log model:

$$log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

   If the model is estimated correctly, then

$$E(y|x) = e^{\sigma^2/2} exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

3. Interpreting level y: implies adjusting for standard error from log models for all the parameters. We can simply use sample standard error $\hat{\sigma}^2$ as proxy for σ^2.

   - ▷ This prediction of level y based on the correction is consistent relying on normally distributed errors
   - ▷ There are no unbiased estimates.

4. **Correct prediction of level without normality of errors**:

$$E(y|x) = \alpha_0 exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

where α_0 is the expected value of $exp(u)$ or $\alpha_0 = E[exp(u)]$. How do we estimated α_0?

(a) **Smearing estimate**: estimated the $log(y)$ model with sample data and then use the OLS residual $\hat{u}_i$ to estimate $\hat{\alpha_0}$ as

$$\hat{\alpha_0} = \frac{1}{n} \sum_{i=0}^{n} exp(\hat{u}_i)$$

(b) **Estimate through origin**: Define $m_i = exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$ and regress m_i on $exp(log(y_i))$ without intercept. The β_j hence obtained is an estimate of α_0 and we denote it by $\check{\alpha_0}$

5. **Goodness of fit with $log(y)$ model**: For the level model estimated as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

R^2 measures squared correlation between y_i and $\hat{y}_i$. But what we have is $\hat{y}_i = \hat{\alpha_0} m_i$. We can use any measure of $\hat{\alpha_0}$ to obtain R^2 because correlations are not affected by change of scale through multiplication.

6. **Bottom line**: use the R^2 compatible with your estimate of $\hat{\alpha_0}$

6.4.4 Bootstrap Standard Errors

1. **Idea of Bootstrapping**: Imagine that you have a sample (large enough), but you have no idea of what the sampling distribution looks like. At times the best proxy of sampling distribution is to treat the big sample as population and draw samples from it.

   (a) Such methods are called re-sampling methods

   (b) Suppose we have an estimate $\hat{\theta}$ which is an estimate of θ. To obtain CIs for $\hat{\theta}$ we do re-sampling called bootstrapping.

2. Bootstrapping original sample of size n by drawing samples with replacement of size n and estimating θ using these different samples. Let us assume we generate m samples. Then the bootstrap standard error of $\hat{\theta}$ is

$$bse(\hat{\theta}) = \Big[\frac{1}{(m-1)}\sum_{b=1}^{m}(\hat{\theta}^b - \bar{\hat{\theta}})^2\Big],$$

   where $\bar{\hat{\theta}}$ is the average of the bootstrap estimates.