Section 10.1 - Simple Linear Regression

Statistics 104

Autumn 2004

Statistical Model for Linear Regression

So far we have only discussed regression as a descriptive technique for bivariate data.

What we have not discussed is what sort of population that the data might have been sampled from and what sort of model could be used to describe the data.

Want to develop a model describing the data generation and which will allow inference on the parameters of that model.

In the examples we've seen before, its possible to have multiple observations at the same x with different y values.

We can think about each x defining a different subpopulation (stratification taken to the extreme) and examining the distribution of the y's for each x.

The linear regression model assumes that for each x, the observed response variable y is normally distributed with a mean that depends on x.

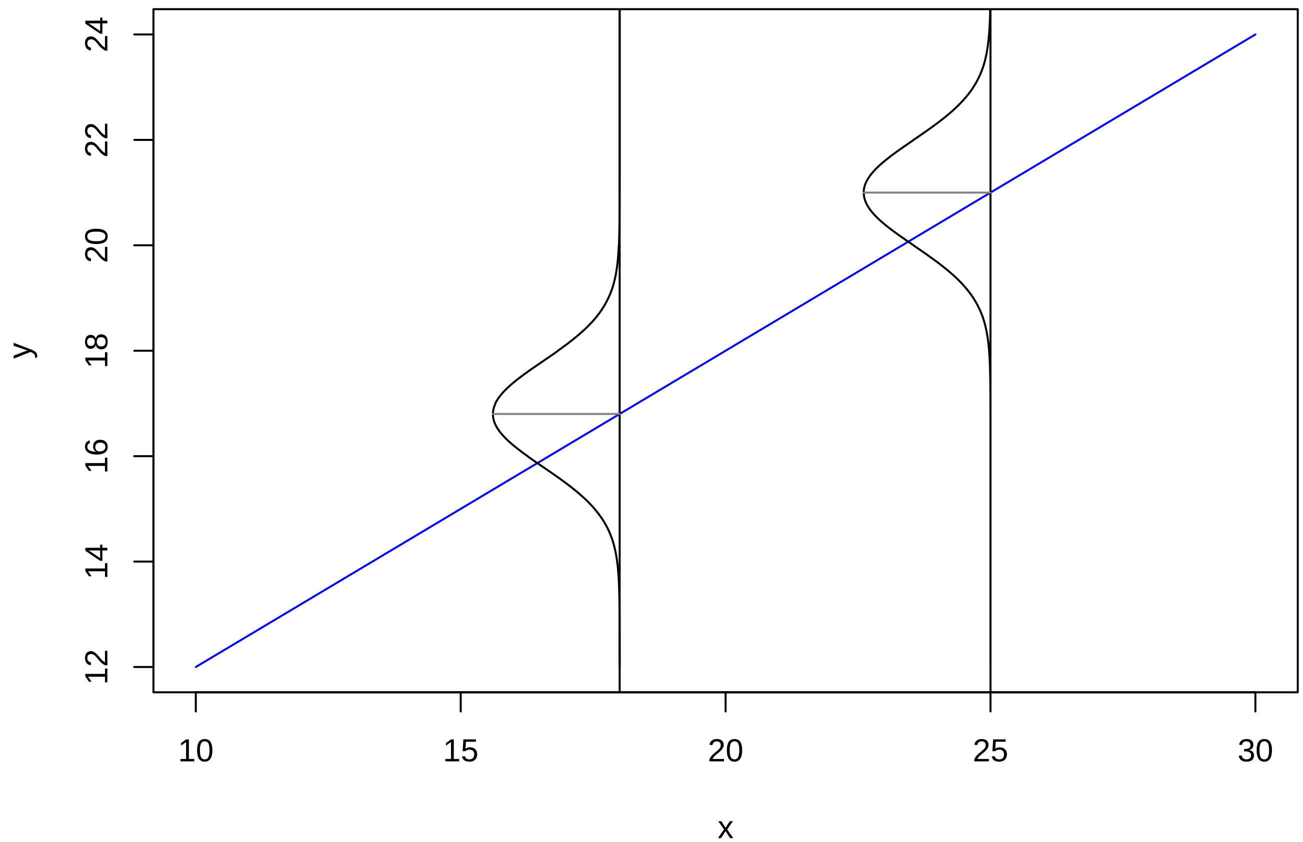
Rather than μ_1 and μ_2 in a two-sample comparision, we are interested on how μ_y changes with x.

In simple linear regression, we assume that the μ_y lie on a line when plotted against x. The equation of the line is

$$\mu_y = \beta_0 + \beta_1 x$$

This is the population regression line.

The observed y's will vary around these means. We will assume that this variation will have the same standard deviation for each x.



When we were discussing regression earlier, we discussed the idea of

$$\text{DATA} = \text{FIT} + \text{RESIDUAL}$$

We can use a similar idea for our population data model

$$\text{DATA} = \text{MEAN} + \text{RANDOM DEVIATION}$$

The Simple Linear Regression model is

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

where the deviations, ϵ_i are assumed to be independent and normally distributed with mean 0 and standard deviation σ ($\epsilon_i \sim N(0, \sigma)$).

The parameters of this model are β_0, β_1, and σ.

Want to address 3 inference problems

1. The slope β_1 and the intercept β_0 of the population regression line.
2. The mean response μ_y for a given value of x.
3. A future response y for a given value of x.

Parameter Estimates

We will continue to use least squares to estimate the parameters.

Recall

$$\begin{aligned} b_1 &= r\frac{s_y}{s_x} \\ b_0 &= \bar{y} - b_1\bar{x} \\ \hat{y} &= b_0 + b_1 x \end{aligned}$$

It can be shown that the sampling distributions of these quantities have means of $\beta_1, \beta_0,$ and μ_y respectively (each is an unbiased estimator).

In addition, each quantity is normally distributed (assuming the deviations are normally distributed).

If they aren't, a more general form of the central limit theorem says they should be approximately normally distributed.

In addition, standard errors for all three quantities can be estimated.

The residuals

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i$$

correspond to the model deviations ϵ_i.

Recall that the e_i's have a sample average of 0, similar to the population mean of the ϵ_i being 0.

We will base our estimate for σ on the e_i's. This is needed to get standard errors for other quantities and may be of interest on its own.

The usual estimate of σ^2 is

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

This is an unbiased estimator of σ^2. In this case s^2 has $n-2$ degrees of freedom.

The usual estimate of σ is

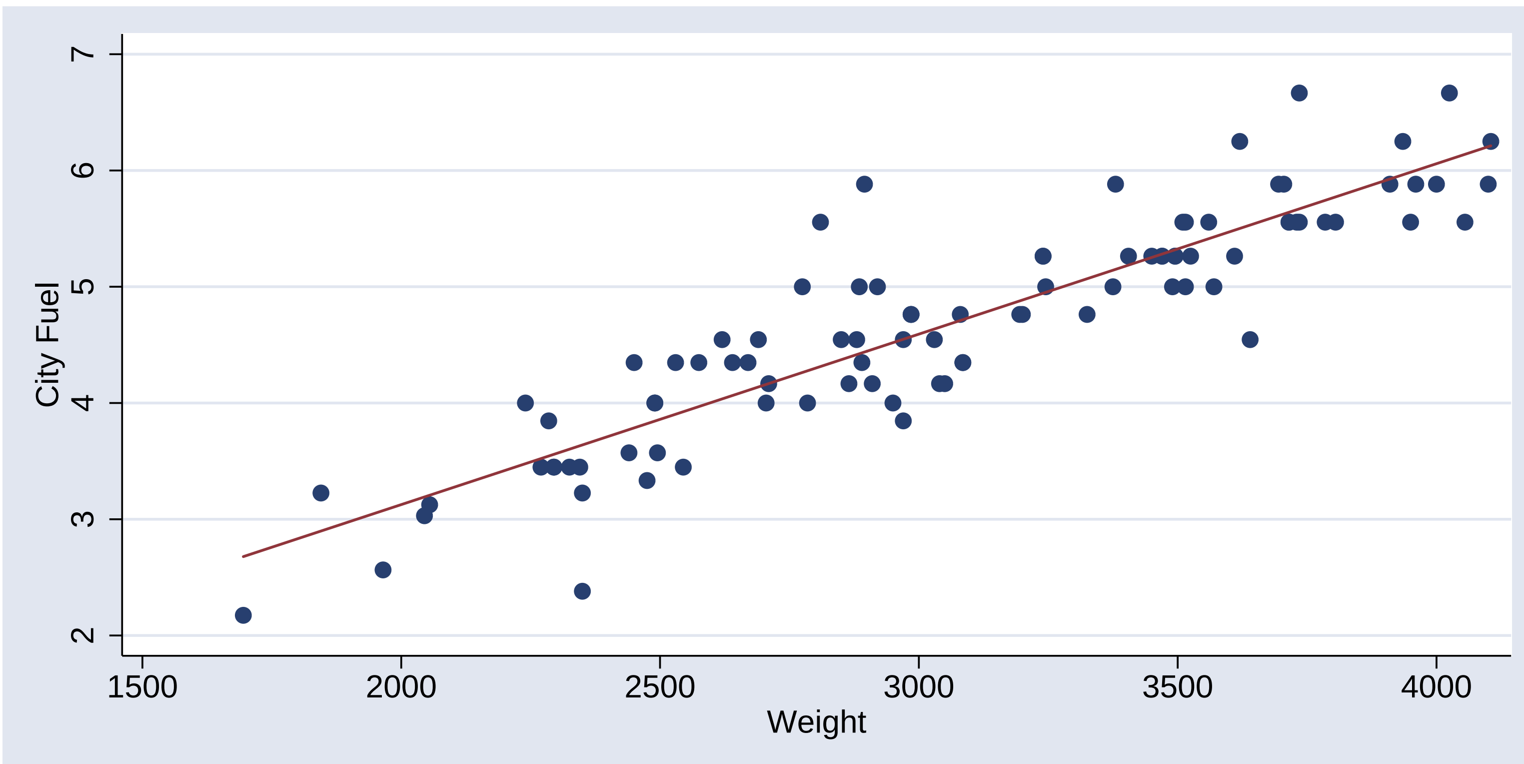
$$s = \sqrt{s^2}$$

As before, we will continue to use a stat package to do the calculations. In particular, the standard errors are difficult to calculate by hand. (We will talk about the formulas for them later.)

Example: City driving fuel use in 1993 cars

$$y = \frac{100}{\text{City MPG}} = \text{City Fuel}$$

This is the number of gallons needed to go 100 miles on average. We want to describe its relationship with car weight.



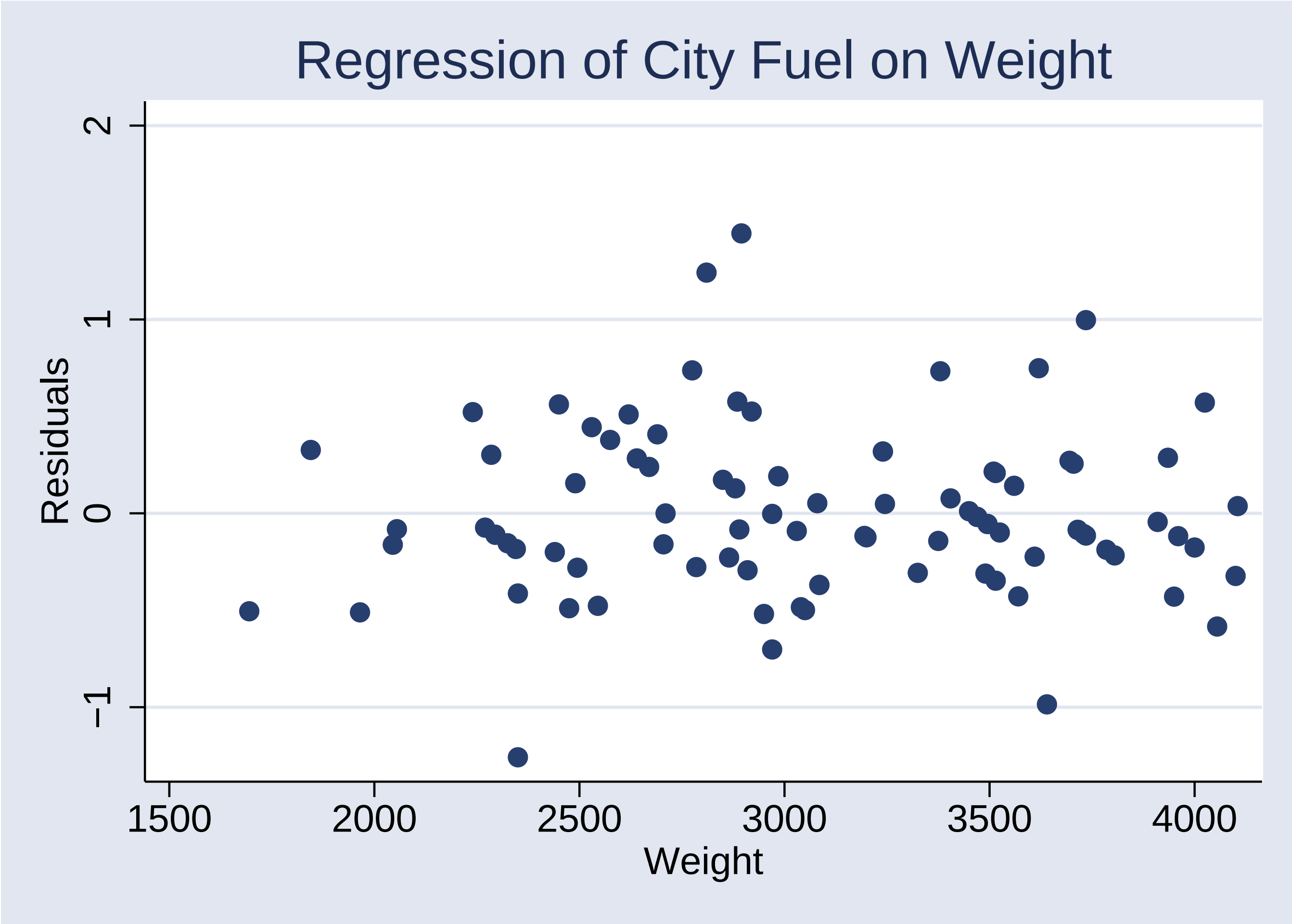
```
. regress cityfuel Weight

  Source |       SS       df       MS              Number of obs =      93
---------+------------------------------           F(  1,    91) =  374.31
   Model |  68.8245208     1  68.8245208           Prob > F      =  0.0000
Residual |  16.7322059    91  .183870394           R-squared     =  0.8044
---------+------------------------------           Adj R-squared =  0.8023
   Total |  85.5567267    92  .929964421           Root MSE      =   .4288

------------------------------------------------------------------------------
cityfuel |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Weight |   .0014662   .0000758     19.35   0.000       .0013157    .0016168
   _cons |   .1936668   .2370884      0.82   0.416      -.2772802    .6646138
------------------------------------------------------------------------------
```

$$b_0 = 0.1937 \qquad s = 0.4288 \qquad \text{(Root MSE)}$$
$$b_1 = 0.001466 \quad s^2 = 0.1839 \quad (\text{MSE } < \text{MS Residual} >)$$

As before, we should check the residual plots



This looks pretty good. There is a suggestion of a couple of outliers, but they don't look too extreme.

Question 1: Inference on β_0 and β_1

As mentioned earlier, b_0 and b_1 are both normally distributed unbiased estimates of β_0 and β_1.

We are in a similar situation as when we are using $\bar{x}$ to estimate μ.

As in that situation, we will use confidence intervals of the form

$$\text{estimate} \pm t^* SE_{\text{estimate}}$$

The confidence intervals are

$$\begin{aligned} \beta_0: &\quad b_0 \pm t^* SE_{b_0} \\ \beta_1: &\quad b_1 \pm t^* SE_{b_1} \end{aligned}$$

where t^* has $n-2$ degrees of freedom and confidence level C.

For 95% confidence intervals, $t^* = 1.986$ $(df = 91 = 93 - 2)$

$$\beta_0 : 0.1937 \pm 1.986 \times 0.2371$$
$$= 0.1937 \pm 0.4709 = (-0.2773, 0.6646)$$

$$\beta_1 : 0.001466 \pm 1.986 \times 0.0000758$$
$$= 0.001466 \pm 0.000151 = (0.001316, 0.001617)$$

```
------------------------------------------------------------------------------
cityfuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Weight |   .0014662   .0000758    19.35   0.000     .0013157    .0016168
   _cons |   .1936668   .2370884     0.82   0.416    -.2772802    .6646138
------------------------------------------------------------------------------
```

Testing on β_0 and β_1 is also similar to case of using $\bar{x}$ to estimate μ. The standard test statistics are:

$$\beta_0 : t = \frac{b_0 - \beta_{0\text{hypoth}}}{SE_{b_0}} \qquad H_0 : \beta_0 = \beta_{0\text{hypoth}}$$

$$\beta_1 : t = \frac{b_1 - \beta_{1\text{hypoth}}}{SE_{b_1}} \qquad H_0 : \beta_1 = \beta_{1\text{hypoth}}$$

Usually the null hypothesis value for both tests is 0.

Note that the test on β_0 is rarely done, as the parameter rarely has great meaning (as we have discussed before).

However the test of whether $\beta_1 = 0$ is often of great interest. If $\beta_1 = 0$ then

$$\mu_y = \beta_0$$

which implies the distribution of y doesn't depend on x in a linear fashion. The t-test on $\beta_1 = 0$ examines whether there is a linear relationship between x and y. This test is usually done two-sided.

For both tests, under the null hypothesis, t have a t distribution with $n-2$ degrees of freedom. There for the p-values for the tests are

$$\begin{array}{ll} H_A : \beta_1 < \beta_{1\text{hypoth}} & p\text{−value} = P[T \leq t_{obs}] \\ H_A : \beta_1 > \beta_{1\text{hypoth}} & p\text{−value} = P[T \geq t_{obs}] \\ H_A : \beta_1 \neq \beta_{1\text{hypoth}} & p\text{−value} = 2 \times P[T \geq |t_{obs}|] \end{array}$$

The p-values are similar for the tests on β_0.

For the example, the tests on whether either of two regression parameters are 0 are

$$\beta_0 : t = \frac{0.1936}{0.2371} = 0.82; \qquad p - \text{value} = 2 \times P[T \geq |0.82|] = 0.416$$

$$\beta_1 : t = \frac{0.001466}{0.0000758} = 19.35; \qquad p - \text{value} = 2 \times P[T \geq |19.35|] \approx 0$$

```
------------------------------------------------------------------------------
cityfuel |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Weight |   .0014662   .0000758    19.35   0.000     .0013157    .0016168
   _cons |   .1936668   .2370884     0.82   0.416    -.2772802    .6646138
------------------------------------------------------------------------------
```

Standard errors for b_0 and b_1

$$SE_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}} = \frac{s}{s_x\sqrt{n-1}}$$

$$SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{s_x^2(n-1)}}$$

Implications of these formulas

1. The less spread out the data is around the regression line (e.g. the smaller s is), the smaller the standard errors.

2. The more data you have, the smaller the standard errors. They both are similar to $\frac{SD}{\sqrt{n}}$.

3. The more spread out your x's (e.g. the bigger s_x is), the more precisely you can estimate the slope.

4. The further you data is centered from 0, the less well you can estimate the intercept.

Question 2: Confidence intervals for a mean response

Interested in the mean response of y when $x = x^*$

$$\mu_y = \beta_0 + \beta_1 x^*$$

Estimate this with

$$\hat{\mu}_y = b_0 + b_1 x^*$$

The confidence interval for μ_y is

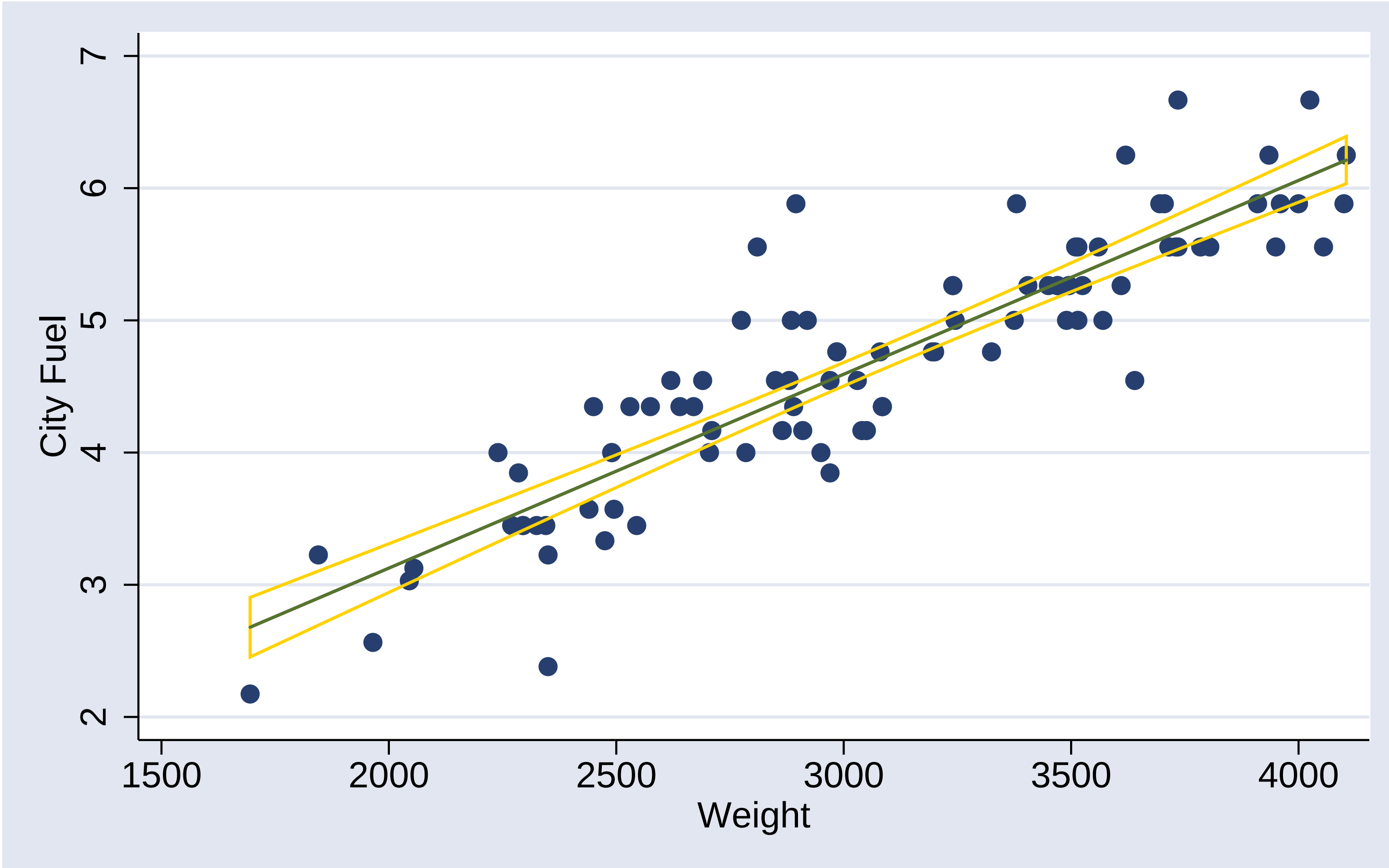
$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}_y}$$

The standard error of $\hat{\mu}_y$ is

$$SE_{\hat{\mu}_y} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{s_x^2(n-1)}}$$

Notice that the SE depends on the x of interest. It is at its smallest when $x^* = \bar{x}$ and increases as x^* moves away from $\bar{x}$.

Also notice that when $x^* = 0$, $\hat{\mu}_y = b_0$ and $SE_{\hat{\mu}_y} = SE_{b_0}$.


City Fuel
7
6
5
4
3
2
1500
2000
2500
3000
3500
4000
Weight

Question 3: Confidence intervals for a future observation (Prediction Intervals)

Interested in a new observation of y when $x = x^*$

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

Estimate this with

$$\hat{y} = b_0 + b_1 x^*$$

The prediction interval for y is

$$\hat{y} \pm t^* SE_{\hat{y}}$$

The standard error of $\hat{y}$ is

$$SE_{\hat{y}} = s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum(x_i - \bar{x})^2}} = \sqrt{s^2 + SE^2_{\hat{\mu}_y}}$$

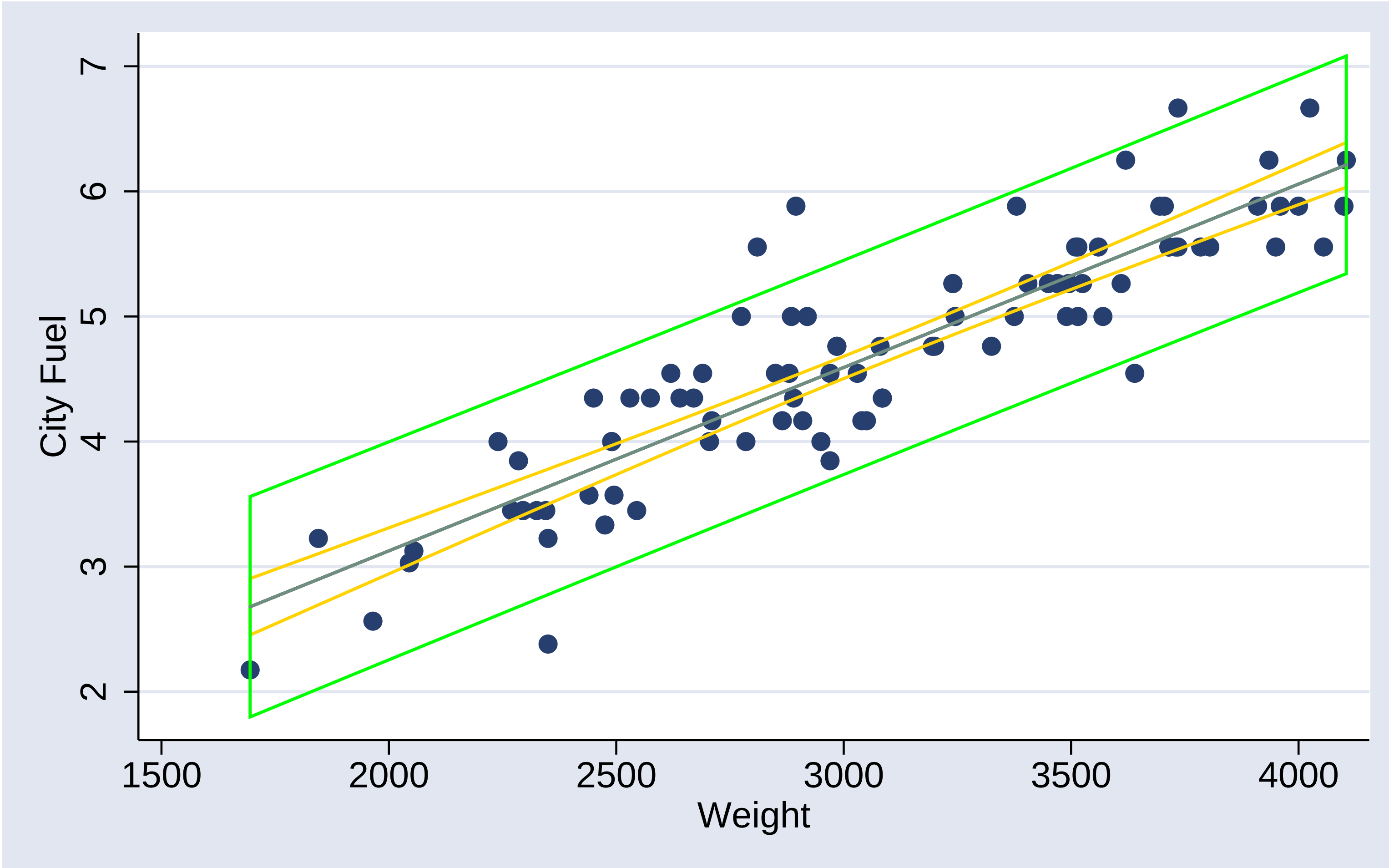
$SE_{\hat{y}}$ deals with 2 pieces of uncertainty

1. Uncertainty about the regression line at x^*

2. Deviations of observations from the true regression line

Notice that $SE_{\hat{y}} \geq SE_{\hat{\mu}_y}$ and $SE_{\hat{y}} \geq s$

Again notice that SE depends on the x of interest. It is at its smallest when $x^* = \bar{x}$ and increases as x^* moves away from $\bar{x}$.



Notice that the prediction interval is wider than the confidence interval for μ_y for every x^*. This is to be expected by the formulas for the standard errors.

Lets compare the 95% CI for μ_y with the 95% Prediction Interval (PI) for y when $x^* = 2000$ and 3000 lbs.

x^*	$\hat{y}$	SE_{μ_y}	$SE_{\hat{y}}$
2000	3.126	0.0927	0.4387
3000	4.592	0.0448	0.4311

95% CI's

$x^* = 2000$: $CI = 3.126 \pm 1.986 \times 0.0927 = 3.126 \pm 0.184$

$x^* = 3000$: $CI = 4.592 \pm 1.986 \times 0.0448 = 4.592 \pm 0.089$

95% PI's

$x^* = 2000$: $CI = 3.126 \pm 1.986 \times 0.4387 = 3.126 \pm 0.871$

$x^* = 3000$: $CI = 4.592 \pm 1.986 \times 0.4311 = 4.592 \pm 0.856$

Notice that the intervals are narrower when $x^* = 3000$ than when $x^* = 2000$ (and PIs are wider than CIs).