M462 Friday March 3, 2006.

D. Levin

Reminder: Regression

Simple Linear Model: $Y_1, \ldots, Y_n$ independent Normals with

$$E(Y_k|x_k) = \beta_0 + \beta_1 x_k$$

$$\text{Var}(Y_k|x_k) = \sigma^2.$$

Parameters: β_0, β_1, σ. Estimators: $\hat{\beta}_0, \hat{\beta}_1, S$.

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

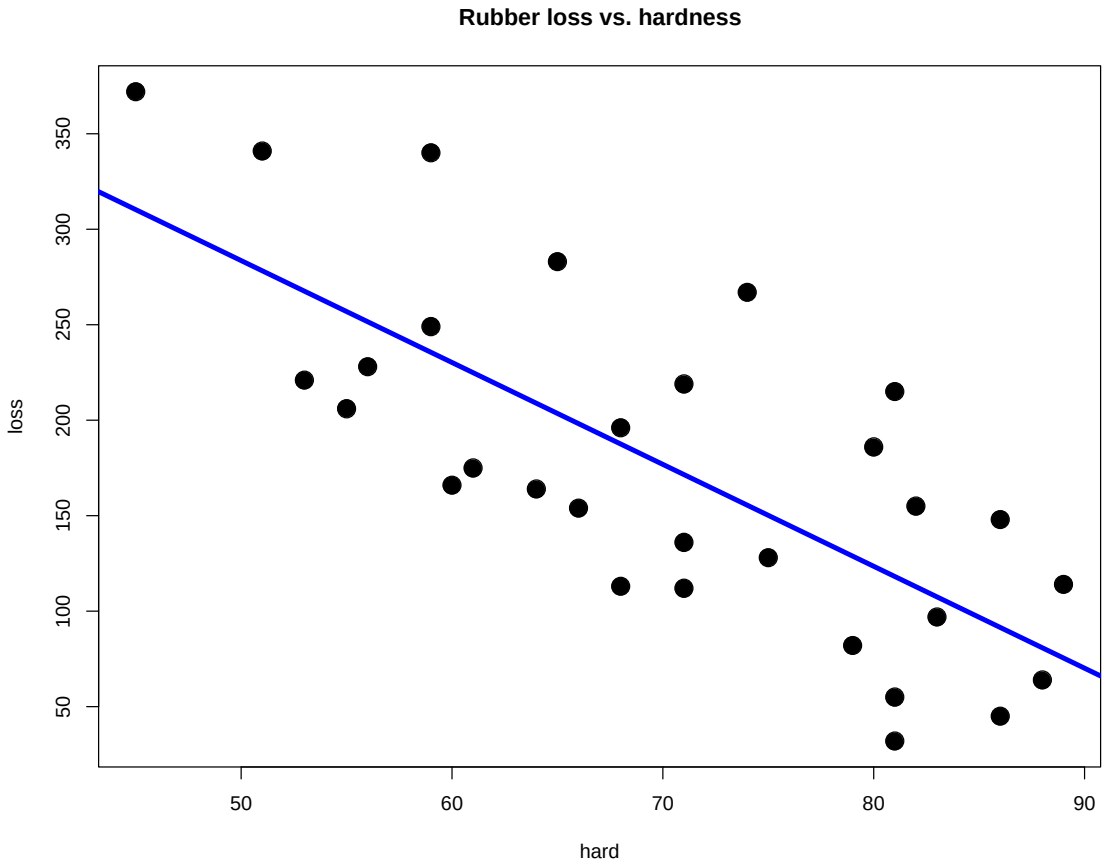
Reminders, continued

- The regression line is (for simple linear model)

$$x \mapsto E(Y|x) = \beta_0 + \beta_1 x. \qquad (\star)$$

- The regression line is estimated by the least-squares line.
- For the simple linear model, the slope and intercept of the least-squares line are the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters β_0 and β_1 in $(\star)$.

Example: Tire rubber



Estimating $E(Y|x)$

Since

$$E(Y|x=60) = \beta_0 + \beta_1 60,$$

and $\hat{\beta}_0$ and $\hat{\beta}_1$ estimate β_0 and β_1,

$$\widehat{E(Y|x=60)} = \hat{\beta}_0 + \hat{\beta}_1 60$$

estimates $E(Y|x=60)$.

The value of the least squares line at x estimates the value of the regression line at x.

Notation: $\hat{Y} := \widehat{E(Y|x)}$ [depends on x]

Inference for $E(Y|x)$

- Goal: Confidence interval for $E(Y|x) = \beta_0 + \beta_1 x$.
  [x fixed.]
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is an unbiased estimator of $E(Y|x)$:

$$E(\hat{Y}) = \beta_0 + \beta_1 x = E(Y|x).$$

- Center of interval then should be $\hat{Y}$; how wide should it be? Depends on the distribution of $\hat{Y}$.

Distribution of $\hat{Y}$

- $$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x = \overline{Y} + (x - \overline{x})\hat{\beta}_1,$$
- $\overline{Y}$ and $\hat{\beta}_1$ are independent Normals,
- Conclusion: $\hat{Y}$ has a Normal distribution.

Standard Error of $\hat{Y}$

The standard deviation of $\hat{Y}$ is

$$\mathrm{SD}(\hat{Y}) = \sqrt{\mathrm{Var}(\overline{Y}) + (x - \overline{x})^2 \, \mathrm{Var}(\hat{\beta}_1)}$$

$$= \sqrt{\frac{\sigma^2}{n} + (x - \overline{x})^2 \, \mathrm{Var}(\hat{\beta}_1)} = \sigma \boxed{\text{stuff}}$$

The *estimated* standard deviation is then

$$\mathrm{SE}(\hat{Y}) = \sqrt{\frac{S^2}{n} + (x - \overline{x})^2 \, \mathrm{SE}(\hat{\beta}_1)^2} = S \boxed{\text{stuff}}$$

Standard errors

If W is a statistic with

$$\mathrm{Var}(W) = \sigma g(x_1, \ldots, x_n),$$

then the standard error of W is the estimated standard deviation of W:

$$\mathrm{SE}(W) = S\sqrt{g(x_1, \ldots, x_n)}.$$

Distribution of $\hat{Y}$

The statistic

$$T = \frac{\frac{\hat{Y}-(\beta_0+\beta_1 x)}{\sigma\ \boxed{\text{stuff}}}}{\sqrt{\frac{(n-2)S^2}{\sigma^2}/(n-2)}} = \frac{\text{standard Normal}}{\sqrt{\text{scaled chi-sq}}}$$

$$= \frac{\hat{Y}-(\beta_0+\beta_1 x)}{S\ \boxed{\text{stuff}}} = \frac{\hat{Y}-(\beta_0+\beta_1 x)}{\text{SE}(\hat{Y})}$$

has a t distribution with $n-2$ deg. of freedom.

Confidence interval for $E(Y|x)$

A $(1-\alpha)\times 100\%$ confidence interval is

$$\hat{Y} \pm t_{\alpha/2,n-2}\,\mathrm{SE}(\hat{Y}).$$

where

$$\mathrm{SE}(\hat{Y}) = \sqrt{\frac{S^2}{n} + (x-\overline{x})^2\,\mathrm{SE}(\hat{\beta}_1)^2}$$

Often S is called residual standard error or just standard error.

Rubber, continued

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 550.4151    65.7867   8.367 4.22e-09 ***
hard         -5.3366     0.9229  -5.782 3.29e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 60.52 on 28 degrees of freedom
Multiple R-Squared: 0.5442,Adjusted R-squared: 0.5279
F-statistic: 33.43 on 1 and 28 DF,  p-value: 3.294e-06
```

$$\hat{Y} = 550.4151 + (-5.34) \times 60 = 230.22$$

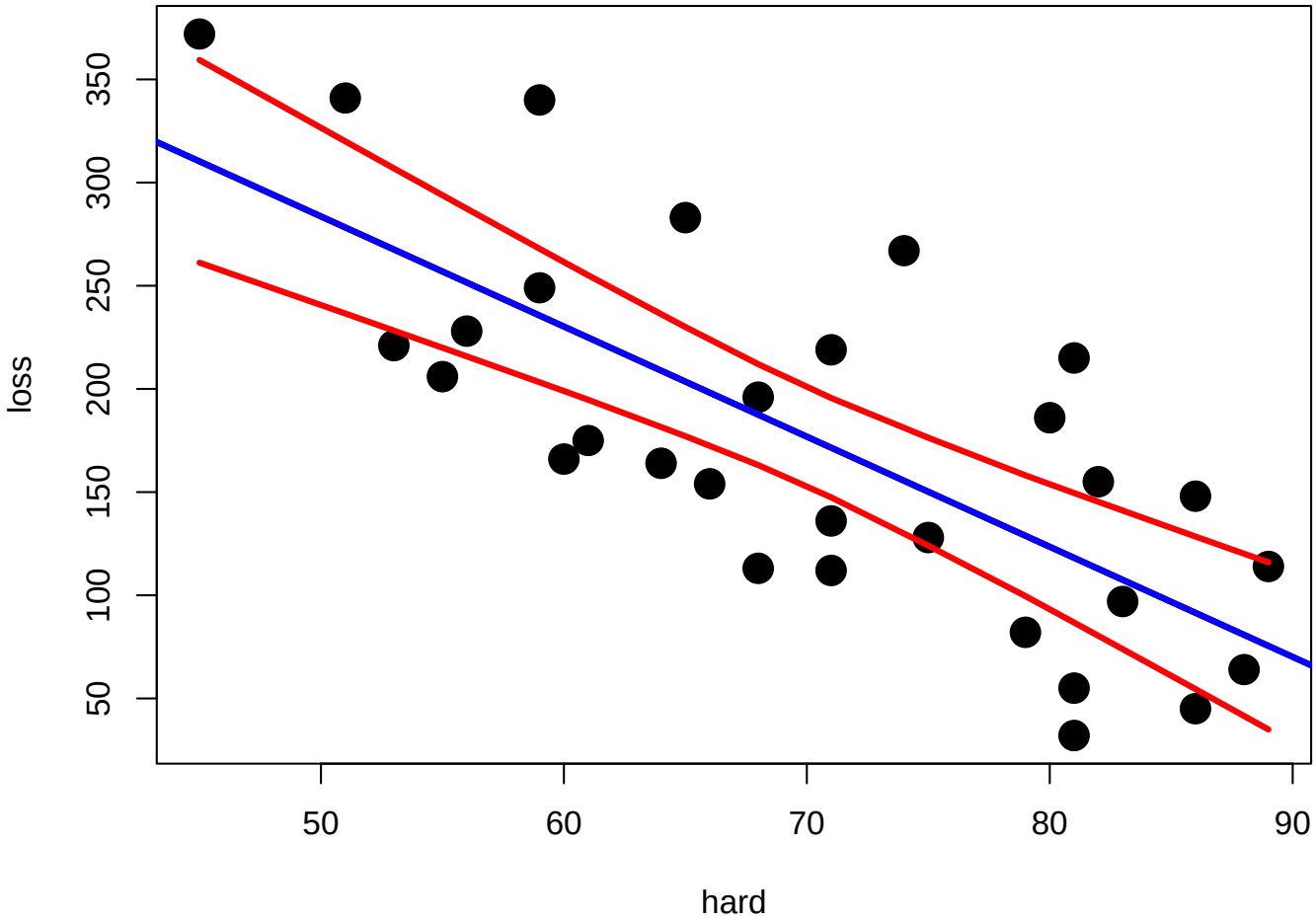
$$SE(\hat{Y}) = \sqrt{\frac{60.52^2}{30} + (60 - 70.27)^2 \times (0.9229)^2} = 14.56.$$

Example, continued

The 95% confidence interval for $E(Y|60)$ is

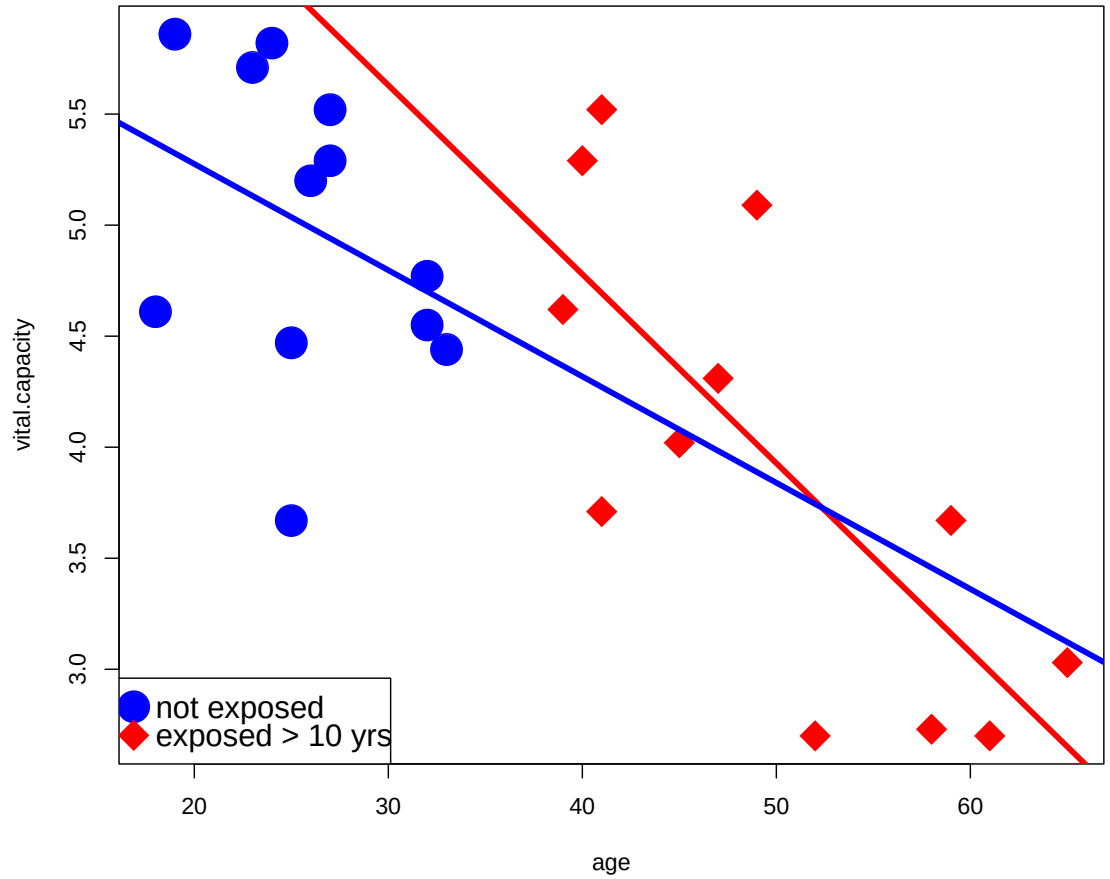
$$230.22 \pm 2.048 \times 14.56 = (200.41,\ 260.04).$$

Confidence bands



Comparing two lines

Data on vital capacity (measure of lung volume) for workers in the cadmium industry.



Comparing models

We have two simple linear models, with two regression lines:

$$E(Y_k|x_k) = \beta_0 + \beta_1 x_k \qquad \text{not exposed}$$

$$E(V_k|u_k) = \gamma_0 + \gamma_1 u_k \qquad \text{exposed .}$$

x and u are ages measured in years, the responses Y, V are vital capacity.

Standard assumptions apply: Each Y_k is Normal with $\text{Var}(Y_k) = \sigma^2$.

a Z-statistic

The statistic

$$Z = \frac{\hat{\beta}_1 - \hat{\gamma}_1 - (\beta_1 - \gamma_1)}{\mathrm{SD}(\hat{\beta}_1 + \hat{\gamma}_1)}$$

has a standard Normal distribution. Note

$$\mathrm{SD}(\hat{\beta}_1 + \hat{\gamma}_1) = \sigma \sqrt{\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2} + \frac{1}{\sum_{i=1}^{n}(u_i - \overline{u})^2}}$$

Pooled estimate of σ

The pooled variance estimate is

$$S_p^2 = \frac{(n-2)S_Y^2 + (m-2)S_V^2}{n+m-4}.$$

The statistic

$$\frac{(n+m-4)S^2}{\sigma^2}$$

has a chi-squared distribution with $n+m-2$ degrees of freedom.

a T-statistic

$$T = \frac{\frac{\hat{\beta}_1 - \hat{\gamma}_1 - (\beta_1 - \gamma_1)}{\mathrm{SD}(\hat{\beta}_1 + \hat{\gamma}_1)}}{\sqrt{\frac{(n+m-4)S_p^2}{\sigma^2} / (n+m-4)}} = \frac{\hat{\beta}_1 - \hat{\gamma}_1 - (\beta_1 - \gamma_1)}{\mathrm{SE}(\hat{\beta}_1 + \hat{\gamma}_1)}$$

has a t distribution with $n+m-2$ degrees of freedom. Here

$$\mathrm{SE}(\hat{\beta}_1 + \hat{\gamma}_1) = S_p \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{1}{\sum_{i=1}^n (u_i - \bar{u})^2}}$$

A test

To test $H_0 : \beta_1 = \gamma_1$ at level α, reject H_0 if and only if

$$\left| \frac{\hat{\beta}_1 - \hat{\gamma}_1}{\mathrm{SE}(\hat{\beta}_1 + \hat{\gamma}_1)} \right| > t_{\alpha/2, n+m-2}.$$

Returning to vital capacity

For the cadmium worker data,

$$t = \frac{-0.04785 - (-0.08511)}{0.04949} = -0.7529.$$

Since $t_{0.025,20} = 2.09$, we do not reject H_0.

Prediction Intervals

We want an interval $[L, U]$, based on our data, so that a future independent observation Y made at x has probability $1 - \alpha$ of falling inside:

$$P(L \leq Y \leq U) = 1 - \alpha.$$

Base interval on

$$Y - \hat{Y}$$

which has a Normal distribution.

Prediction, continued

Need mean and variance of $Y - \hat{Y}$:

$$E(Y - \hat{Y}) = 0$$

$$\text{Var}(Y - \hat{Y}) = \sigma^2 + \text{Var}(\hat{Y})$$

$$\text{SE}(Y - \hat{Y}) = \sqrt{S^2 + \text{SE}(\hat{Y})^2}$$

Prediction, continued

The statistic

$$T = \frac{Y - \hat{Y}}{\mathrm{SE}(Y - \hat{Y})}$$

has a t distribution with $n-2$ deg. of freedom. So

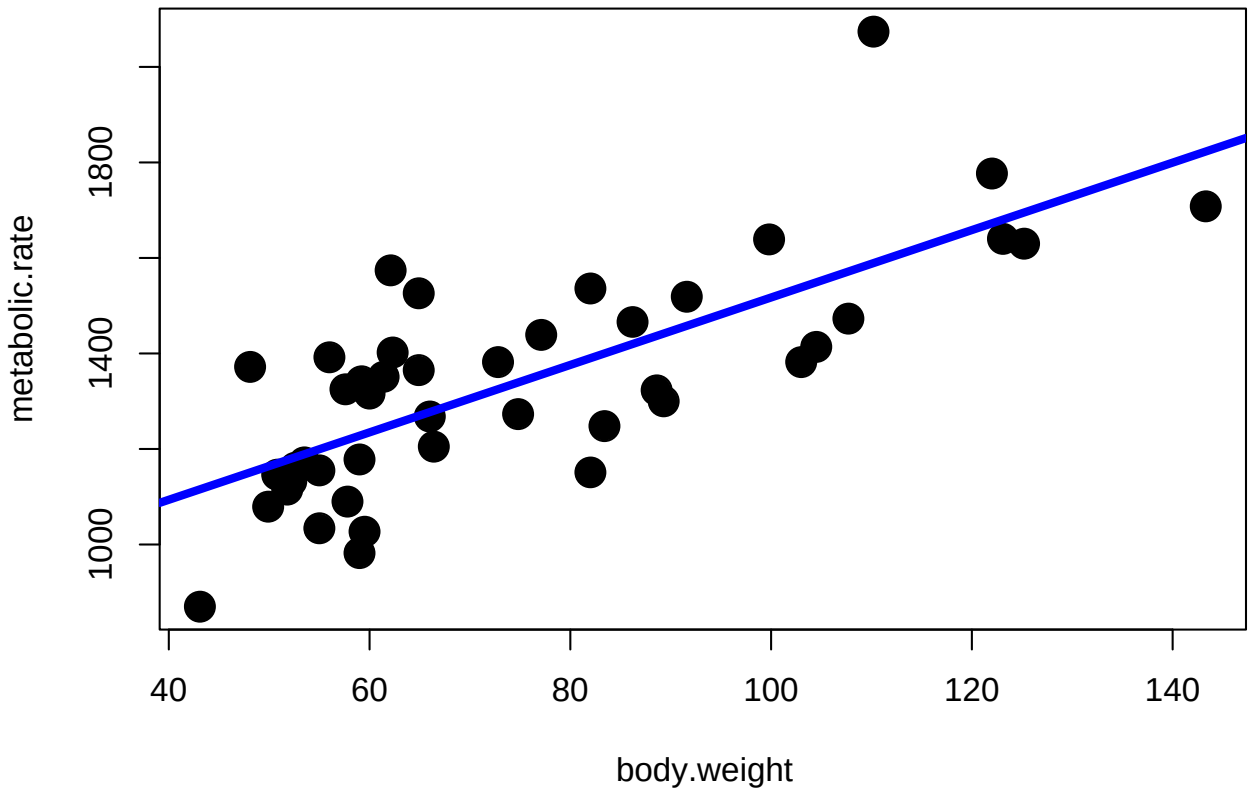
$$\hat{y} \pm t_{\alpha/2,n-2}\,\mathrm{SE}(Y - \hat{Y})$$

is an $(1-\alpha) \times 100\%$ prediction interval.

Prediction, continued

$$
\begin{aligned}
1-\alpha &= P\left(-t_{\alpha/2,n-2} \leq \frac{Y-\hat{Y}}{\mathrm{SE}(Y-\hat{Y})} \leq t_{\alpha/2,n-2}\right) \\
&= P\left(\hat{Y}-t_{\alpha/2,n-2}\,\mathrm{SE} \leq Y \leq \hat{Y}-t_{\alpha/2,n-2}\,\mathrm{SE}\right)
\end{aligned}
$$

Metabolic rate vs. weight



95% prediction interval

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 811.2267    76.9755  10.539 2.29e-13 ***
body.weight   7.0595     0.9776   7.221 7.03e-09 ***
---
Residual standard error: 157.9 on 42 degrees of freedom
```

$$\hat{Y} = 811.2 + 7.06 \times 80 = 1376$$

$$\mathrm{SE}(\hat{Y} - Y) = \sqrt{\left(\frac{157.9^2}{44} + (\underbrace{80 - 74.9}_{80-\bar{x}})^2 0.98^2\right) + 157.9^2}$$

$$= 159.8$$

Prediction bands