Testing hypotheses: Additionally assume that ϵ_i are normal, i.e., $\epsilon_i \sim$ independent $N(0, \sigma^2)$.

- Each $\hat{\beta}_j$ follows a normal distribution
- $(n-k-1)\hat{\sigma}^2/\sigma^2$ follows a χ^2_{n-k-1} distribution

> indep } Proofs in advanced classes.

$E[Y|X] = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$

Testing significance of jth predictor:

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$$

$\Rightarrow$ Here: keep all the other predictors in model.

$\Rightarrow$ check whether X_j is useful or not after adjusting for the effects of the other predictors in model.

- H_0 : Predictor X_j is not useful for predicting response **after adjusting for the other predictors**
- Test statistic: $(\hat{\beta}_j - 0)/\widehat{SE}(\hat{\beta}_j)$
- Null distribution: $\sim$ t error d.f. when H_0 is true.
- Rejection region:
- p-value:
- $100(1-\alpha)\%$ CI for β_j:

— just a t-test and recall how to do this.

Testing model significance:

$$H_0 : \beta_1 = \ldots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \text{at least one } \beta_j \neq 0$$

- H_0 : None of the predictors is useful for predicting response
- Test statistic: $\frac{MS_{reg}}{MS_{error}} \sim F_{k, n-k-1}$ if H_0 is true.
- Null distribution: k = # slopes; $n-k-1$ = error d.f.
- Rejection region:
- p-value:

Know how to do this.

→ an F-test.

Coefficient of determination: As before,

$$R^2 = \frac{SS_{\text{REG}}}{SS_{\text{TOT}}} = 1 - \frac{SS_{\text{ERR}}}{SS_{\text{TOT}}}$$

- R^2 = proportion of total variation explained by regression
- R^2 = square of correlation between $(Y_i, \hat{Y}_i)$, $i = 1, \ldots, n$ — can be verified
- When we add new predictors to the model, R^2 ↑.
- Even adding useless predictors will also ↑ R^2.
- Not a fair criterion for comparing models with different numbers of predictors

Beware of overfitting a model:

- Overfitting = Having too many predictors in the model
- An overfitted model will provide a good fit to the data at hand, but it will be terrible at predicting future observations.
- Estimated regression coefficients have large standard errors.
- See Figure 11.4 on page 369

⇓

estimates are unreliable

⇓

predictions are not reliable in the sense that they will have large SE's.

Adjusted R^2:

$$R^2_{\text{adj}} = 1 - \frac{SS_{\text{ERR}}/(n-k-1)}{SS_{\text{TOT}}/(n-1)}$$

- Unlike R^2, R^2_{adj} rewards adding a predictor only if it reduces the error SS considerably
- Imagine adding a useless predictor. In this case, SS_{REG} and hence SS_{ERR} does not change much. However, $SS_{\text{ERR}}/(n-k-1)$ increases, which in turn decreases R^2_{adj}
- A more fair measure of goodness-of-it than R^2
- Can be used to compare two models with different numbers of predictors — choose the model with the highest R^2_{adj}

Comparing two Nested Models

Nested models: Model 2 is nested within Model 1 if the predictors of Model 2 are a subset of predictors of Model 1.

Issue: How to compare two nested models?

Full model: Predictors $X_1, \ldots, X_m$

Reduced model: Predictors $X_1, \ldots, X_k$, i.e., it does not have predictors $X_{k+1}, \ldots, X_m$

Hypotheses: $H_0 : \beta_{k+1} = \ldots = \beta_m = 0$, vs., H_1 : at least one slope $\neq 0$

Extra sum of squares: Difference in variation explained by the two models

$$\begin{aligned} SS_{\text{EX}} &= SS_{\text{REG}}(\text{full}) - SS_{\text{REG}}(\text{reduced}) \\ &= SS_{\text{ERR}}(\text{reduced}) - SS_{\text{ERR}}(\text{full}) \end{aligned}$$

regression slopes set to zero under H_0.

- SS_{EX} has $m-k$ degrees of freedom. It equals the number of regression coefficients set to zero under H_0.

Test statistic:

$\sim F_{m-k,\, n-m-1}$ if H_0 is true.

$$F = \frac{MS_{EX}}{MS_{\text{ERR}}(\text{full})} = \frac{SS_{EX}/(m-k)}{SS_{\text{ERR}}(\text{full})/(n-m-1)}$$

↑ NOTE: THIS IS NOT MS_{ERR}(reduced).

Null distribution:

Rejection region:

p-value:

} F-test.

- aka "partial F-test"
- Used for designing stepwise model selection procedures (see pages 392-394)
- To test $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$ — can do partial F test (with num d.f. 1 and denom d.f. = err d.f.)
  can see that: $F = t_j^2$ — two tests are equivalents (their p-values are the same)

Example: Home price data. These data come from a sample of homes sold in Maplewood, NJ in 2001.

```
# Read the home price data

home <- read.table("homeprice_multiple_predictors.txt",
sep=",", header=T)
```

```
> str(home)
'data.frame': 29 obs. of  7 variables:
 $ list        : num  80 151 310 295 339 ...
 $ sale        : num  118 151 300 275 340 ...
 $ full        : int  1 1 2 2 2 1 3 1 1 1 ...
 $ half        : int  0 0 1 1 0 1 0 1 2 0 ...
 $ bedrooms    : int  3 4 4 4 3 4 3 3 3 1 ...
 $ rooms       : int  6 7 9 8 7 8 7 7 7 3 ...
 $ neighborhood: int  1 1 3 3 4 3 2 2 3 2 ...
>
```

```
# Attach the dataset in R's memory so that we can
# directly use the names of the variables

attach(home)

# Look at distributions of some predictors

> table(bedrooms)
bedrooms
 1  2  3  4  5
 1  3 16  8  1
>
> table(full)
full
 1  2  3
13 11  5
>
```

First step: Exploratory analysis (boxplots, scatterplots, etc.)

```
> table(half)
half
 0  1  2
13 13  3
>
> table(neighborhood)
neighborhood
 1  2  3  4  5
 2  8 12  5  2
>

# Regress sale price on # bedrooms and neighborhood

fit1 <- lm(sale ~ bedrooms + neighborhood)

> summary(fit1)

Call:
```

```
lm(formula = sale ~ bedrooms + neighborhood)

Residuals:
    Min      1Q  Median      3Q     Max
-90.871 -39.861   0.636  28.815 107.660

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -132.057     40.341  -3.273 0.003001 **
bedrooms      42.483     11.446   3.712 0.000987 ***
neighborhood  93.493      9.101  10.273 1.21e-10 ***
---
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1

Residual standard error: 47.3 on 26 degrees of freedom
Multiple R-squared: 0.8491,Adjusted R-squared: 0.8375
F-statistic: 73.16 on 2 and 26 DF,  p-value: 2.1e-11
>
```

keeping all the remaining predictors in the model.

$\hat{\beta}/\widehat{SE}$

$\hat{\beta}$ $\widehat{SE}$ p-value

β_0 β_1 β_2

$H_0: \beta_0 = 0$ vs. $H_1: \beta_0 \neq 0$

$H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$

$H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$

$\hat{\sigma}$ error d.f.

$H_0: \beta_1 = \beta_2 = 0$ vs. H_1: At least one is not zero.

```
# Add # full and half baths

fit2 <- update(fit1, . ~ . + full + half)

> summary(fit2)

Call:
lm(formula = sale ~ bedrooms + neighborhood + full + half)

Residuals:
    Min      1Q  Median      3Q     Max
-56.554 -38.067   6.027  26.998  53.311

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -125.121     33.136  -3.776 0.000926 ***
bedrooms       29.513     10.091   2.925 0.007419 **
```

Either 'lm' ~ 'update'

```
neighborhood   78.724      9.669    8.142 2.31e-08 ***
full           27.345     13.604    2.010 0.055785 .
half           45.553     12.129    3.756 0.000974 ***
---
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1

Residual standard error: 38.79 on 24 degrees of freedom
Multiple R-squared: 0.9063,Adjusted R-squared: 0.8907
F-statistic: 58.05 on 4 and 24 DF,  p-value: 5.425e-12
>

# Drop # full baths

fit3 <- update(fit2, . ~ . - full)

> summary(fit3)

Call:
```

Partial F-test for $\beta_{full} = 0$ will have p-value = 0.055785

```
lm(formula = sale ~ bedrooms + neighborhood + half)

Residuals:
    Min      1Q  Median      3Q     Max
-67.55 -42.27   7.17  26.93  68.83

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -127.348     35.073  -3.631  0.00127 **
bedrooms       35.649     10.187   3.500  0.00177 **
neighborhood   90.982      7.947  11.449 1.95e-11 ***
half           37.004     12.030   3.076  0.00503 **
---
Signif. codes: 0‘***’0.001‘**’0.01‘*’0.05‘.’0.1‘ ’1

Residual standard error: 41.08 on 25 degrees of freedom
Multiple R-squared: 0.8905,Adjusted R-squared: 0.8774
F-statistic:  67.8 on 3 and 25 DF,  p-value: 3.808e-12
```

```
>

# Compare the nested models

Check {\tt ?anova.lm}
```

Important note: When comparing two models using anova the results are as expected from the partial F-test. However, when more than two models are compared using anova, the F-statistic and p-value may not be what we would like. The reason for this is that the F-statistic compares the mean SS for a row to the MS_{ERR} for the **largest model** considered.

```
> anova(fit1, fit3, fit2)
```

Be Careful in interpretation.

```
Analysis of Variance Table

Model 1: sale ~ bedrooms + neighborhood
Model 2: sale ~ bedrooms + neighborhood + half
Model 3: sale ~ bedrooms + neighborhood + full + half
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     26 58164
2     25 42194  1   15970.1 10.6132 0.003338 **
3     24 36114  1    6080.1  4.0406 0.055785 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
>

> anova(fit1, fit2)
Analysis of Variance Table

Model 1: sale ~ bedrooms + neighborhood
```

reduced model

$H_0: \beta_{full} = \beta_{half} = 0$ vs H_1: At least one is $\neq 0$.

```
Model 2: sale ~ bedrooms + neighborhood + full + half
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     26 58164
2     24 36114  2     22050 7.3269 0.003283 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
>
```

Full model

F

p-val of

dfnm d.f.

dfnm d.f.

SS_{EX}

$(SS_{EX}/q) / MS_{ERR}(full)$

```
> anova(fit3, fit2)
Analysis of Variance Table
```

reduced

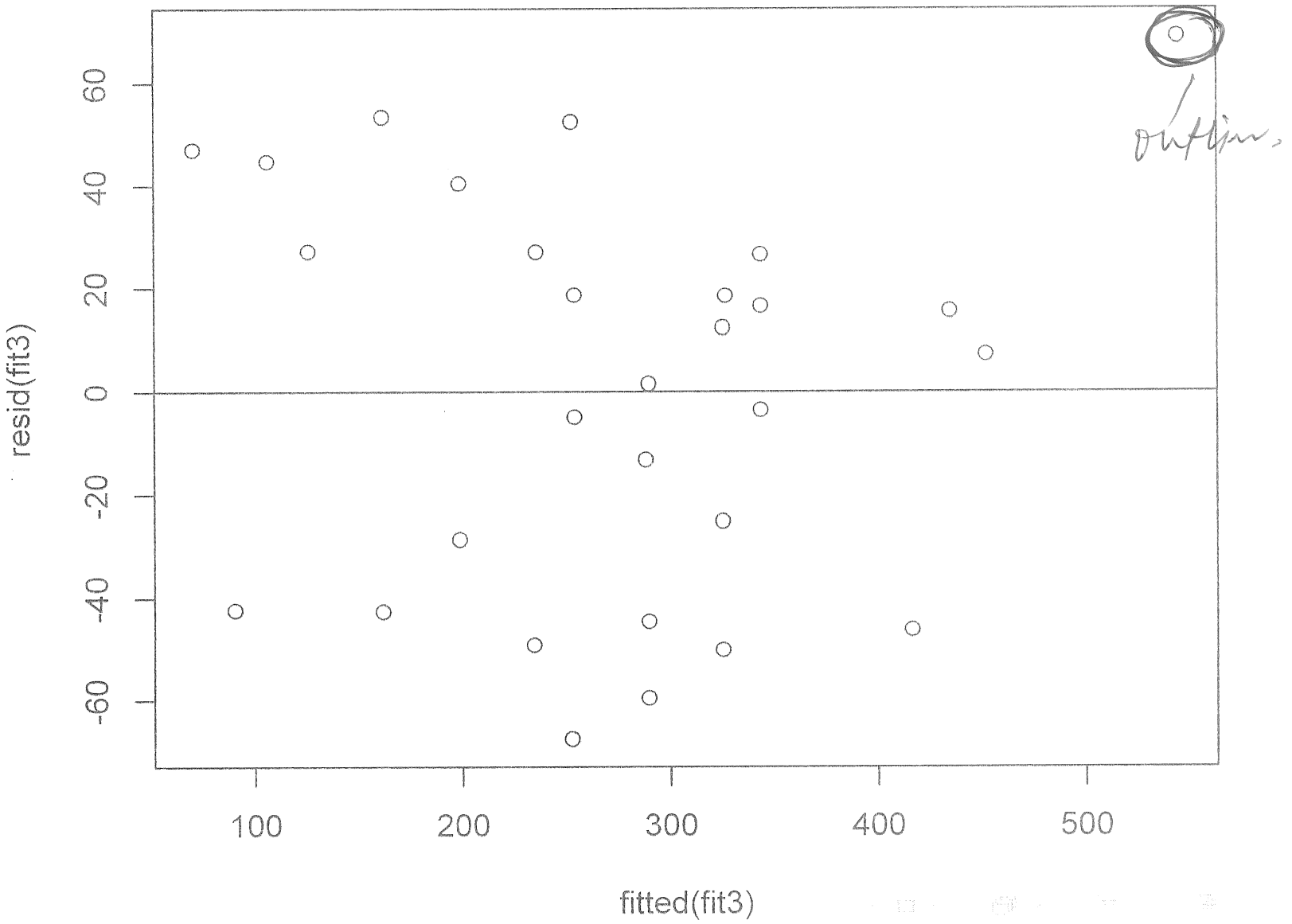
```
Model 1: sale ~ bedrooms + neighborhood + half
Model 2: sale ~ bedrooms + neighborhood + full + half
  Res.Df   RSS Df Sum of Sq      F  Pr(>F)
1     25 42194
2     24 36114  1    6080.1 4.0406 0.05579 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

full model

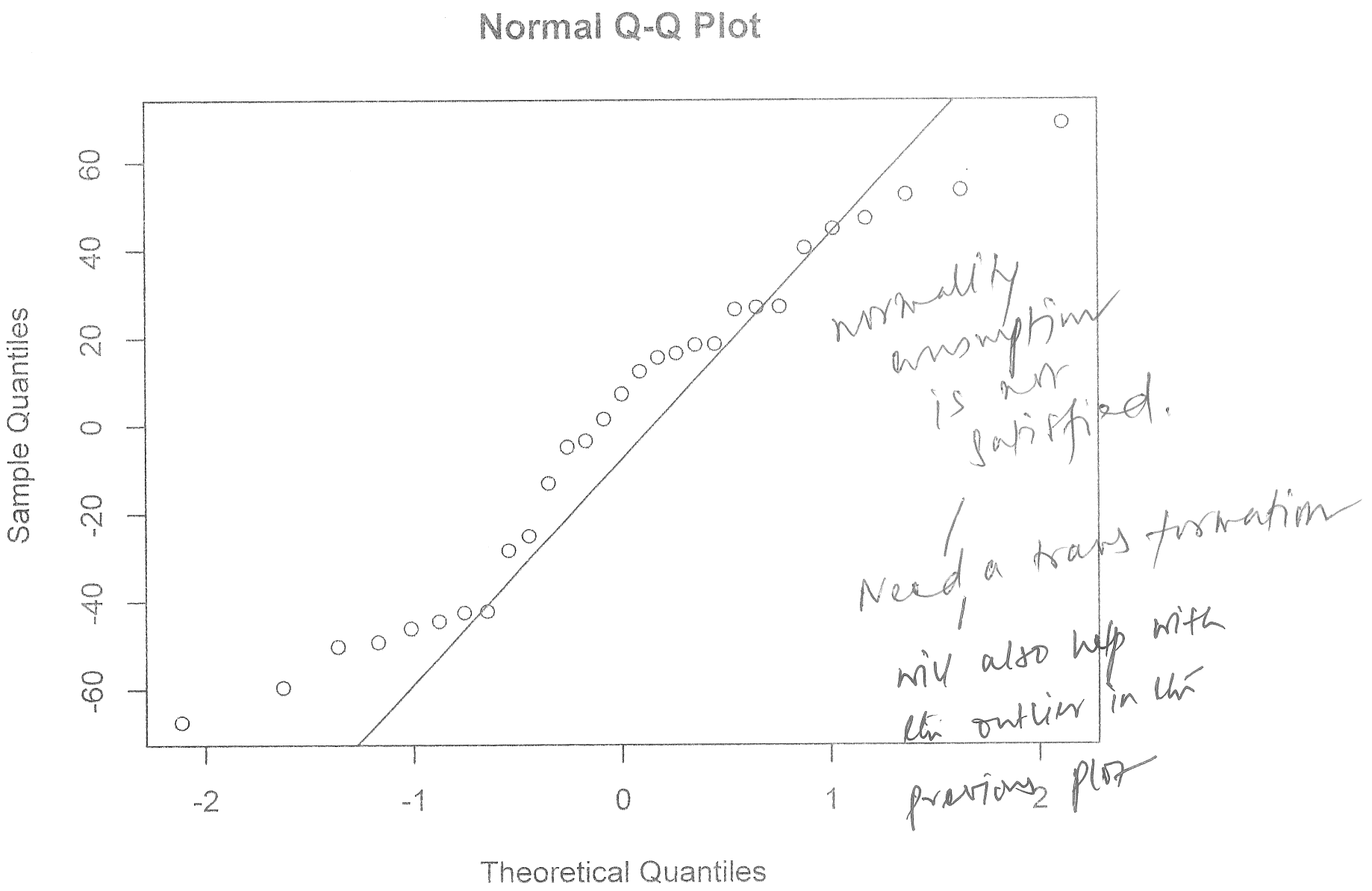
same as the p-value for the t-test.

```
>

# Residual plot

plot(fitted(fit3), resid(fit3))
abline(h=0)

# QQ plot

qqnorm(resid(fit3))
qqline(resid(fit3))
```

Residual plot



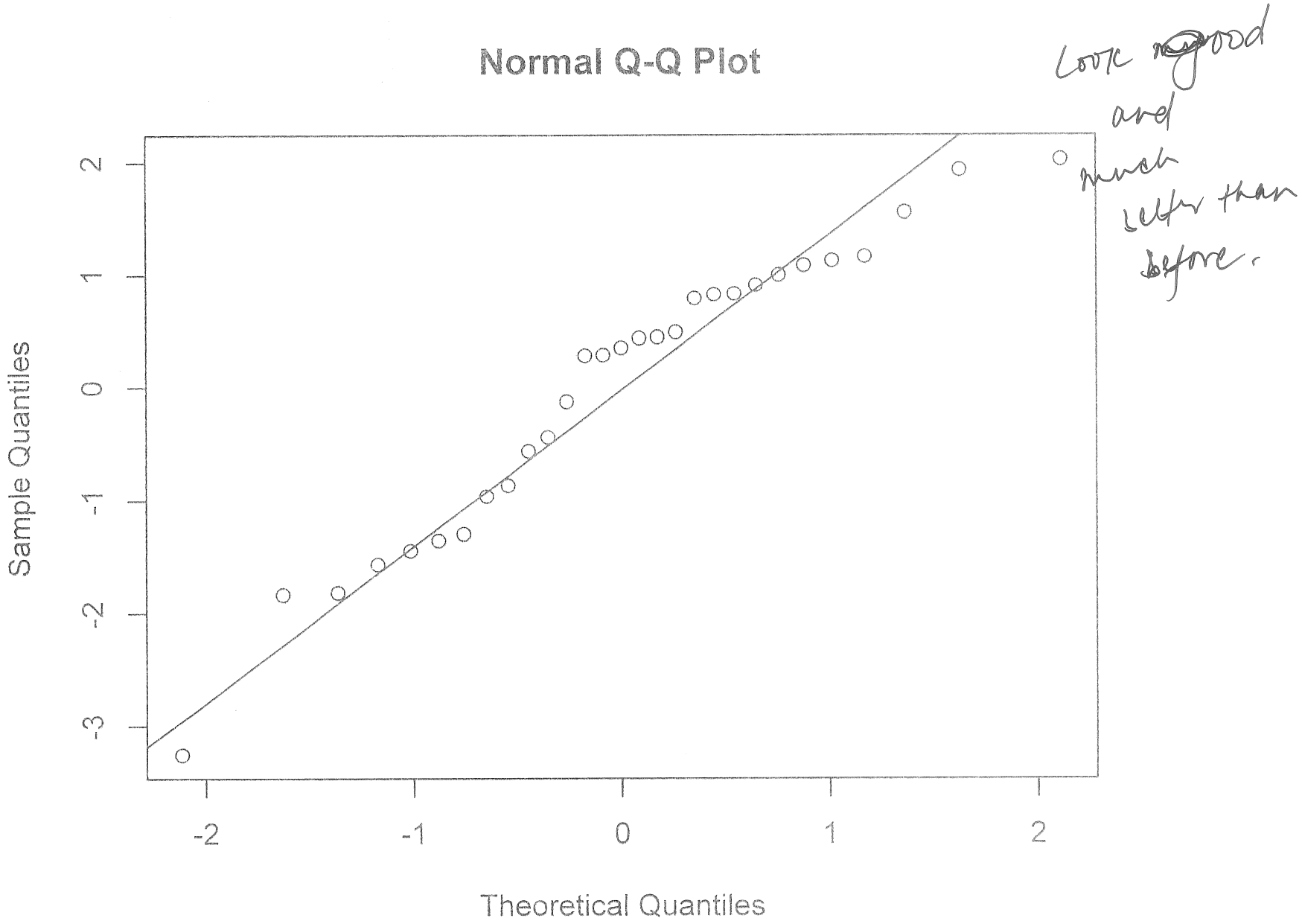
Normal QQ plot



```
# Take sqrt(sale) rather than sale as response

fit4 <- update(fit3, sqrt(sale) ~ .)

# New QQ plot

qqnorm(resid(fit4))
qqline(resid(fit4))
```

sqrt transformation of sale

Normal QQ plot for transformed data



Look much good and much better than before.